

# Decadal climate predictions using sequential learning algorithms

Ehud Strobach<sup>1</sup> and Golan Bel<sup>1</sup>

<sup>1</sup>*Department of Solar Energy and Environmental Physics,*

*Blaustein Institutes for Desert Research,*

*Ben-Gurion University of the Negev, Sede Boqer Campus 84990, Israel*

## Abstract

Ensembles of climate models are commonly used to improve climate predictions and assess the uncertainties associated with them. Weighting the models according to their performances holds the promise of further improving their predictions. Here, we use an ensemble of decadal climate predictions to demonstrate the ability of sequential learning algorithms (SLAs) to reduce the forecast errors and reduce the uncertainties. Three different SLAs are considered, and their performances are compared with those of an equally weighted ensemble, a linear regression and the climatology. Predictions of four different variables—the surface temperature, the zonal and meridional wind, and pressure—are considered. The spatial distributions of the performances are presented, and the statistical significance of the improvements achieved by the SLAs is tested. Based on the performances of the SLAs, we propose one to be highly suitable for the improvement of decadal climate predictions.

## I. INTRODUCTION

Global circulation models are the main tools used to simulate future climate conditions. There are two main practices by which to initialize these models that represent predictions for two different time scales. The first practice corresponds to long-term climate projections. In this type of simulation, the climate models are initialized in the pre-industrial era (aka uninitialized runs) and integrated forward in time (usually until 2100). In these simulations, the atmospheric composition in the past is set according to observations, while for the future, several representative concentration pathways [1], corresponding to different scenarios of atmospheric composition changes, are used. These climate simulations are expected to provide information about the response of the climate system to different emission scenarios by predicting the changes in the long-term averages (10 years and more) and the statistics of climate variables, under different atmospheric composition scenarios [2].

The second practice, which is considered in this work, is near-term (decadal) climate predictions intended to provide information on the dynamics of the climate system in time scales shorter than those of significant changes in the atmospheric concentration and the response time of the climate system to such changes. In this practice, the climate models are initialized with observed conditions close to the prediction period. The expected information from these simulations is the dynamics of the monthly to decadal averages of climate variables [3–6], which is of great importance for climate services [7]. Recent studies have demonstrated a potential decadal prediction skill in different regions and for different physical processes [4, 5, 8–10].

Despite their relatively short term, decadal climate predictions are still accompanied by large uncertainties, and new methods to improve the predictions and reduce the associated uncertainties are of great interest. One of the main approaches to improving climate predictions is to combine the output from an ensemble of climate models. This approach has two known advantages compared with single model predictions. First, it was shown that the ensemble average generates improved predictions [11–16]; second, the distribution of the ensemble member predictions can provide an estimate of the uncertainties. However, the simple average of climate simulations does not account for the quality differences between the ensemble members; therefore, it is expected that weighting the ensemble members based on their past performances will increase the forecast skill.

Uncertainties in climate predictions can be attributed to three main sources. The first is internal

variability, that is, uncertainties due to different initial conditions (either different initialization times or different initialization methods) that were used to run a specific model. The second source is model uncertainties due to different predictions of different models. The third source is forcing scenario uncertainties due to different scenarios assumed for the future atmospheric composition [17]. The contribution of these sources to the total uncertainty of the climate system varies with the prediction lead time and is also spatially, seasonally and averaging-period dependent [18]. It was shown that for global and regional decadal climate predictions, scenario uncertainties are negligible compared to the first two sources [17, 19].

There are two contributions to the internal variability—variability due to different starting conditions and variability due to different initialization methods. Uncertainties due to different starting conditions stem from the chaotic nature of the simulated climate dynamics and cannot be reduced using the ensemble approach. However, uncertainties due to different initialization methods and the model variability can be reduced by weighting the members of the ensemble. The total reduction of the uncertainty depends on the relative contribution of these sources to the total uncertainty.

Bayesian inference is one of the methods that have been used in the past to weight an ensemble of climate models. The main part of this method is the calculation of the posterior density, which is proportional to the product of the prior and the likelihood. The Bayesian method optimizes the probability density function (PDF) of the climate variable to the PDF of the data during a learning period and uses it for future predictions. It does not assign weights to the climate models; instead, it gives an estimation for the PDF of the predicted climate variable. Bayesian inference has been used extensively for projections of future climate [20–28] and also for near-term climate predictions [29, 30]. The use of Bayesian inference has reduced the uncertainties of the climate projections and improved their near-term predictions. However, this method relies on many assumptions regarding the distribution of the climate variables that are not always valid, making the Bayesian inference subjective and variable-dependent.

A second, and more common, method that has been used to improve climate predictions is linear regression [31–42]. The linear regression method does not assign weights to the ensemble members but rather attempts to find a set of coefficients such that the scalar product of the vector of coefficients and the vector of the model predictions yields the minimal sum of squared errors relative to past observations. The same set of coefficients is then used to produce future predictions. As a consequence, the regression can be used only for deterministic predictions, that is, the linear combination of the models is calculated to produce better predictions, but there is no

straightforward method to estimate the associated uncertainties. Similarly to the Bayesian method, the regression method also relies on a few inherent assumptions, such as the normal distribution of the prediction errors (therefore, defining the optimal coefficients as those minimizing the sum of squared errors) and the independence of the ensemble member predictions.

Sequential learning algorithms (SLAs, also known as online learning) [43] weight ensemble members based on their past performances. These algorithms were shown to improve long-term climate predictions [44, 45] and seasonal to annual ozone concentration forecasts [46, 47]. More recently, it was shown that decadal climate predictions of the 2m-temperature can be improved using SLAs and can even become skillful when the climatology is added as a member of the ensemble [48]. The SLAs have several advantages over the other ensemble methods described above. First, they do not rely on any assumption regarding the models and the distribution of the climate variables. In addition, the weights assigned to the models can be used for model evaluation and the comparison of different parameterization schemes or initialization methods. Third, the weighted ensemble provides not only predictions but also the associated uncertainties. All these characteristics suggest that the SLAs are suitable for the improvement of various climate variable predictions.

Here, we test the performances of SLAs in predicting the, previously investigated, 2m-temperature and three additional climate variables—namely, the zonal and meridional components of the surface wind and the surface pressure. The results of the CMIP5 [49] decadal experiments constitute the ensemble, and the NCEP reanalysis data [50] are considered as the observations. The performances of the SLAs are compared with those of the regression method. The comparison with the Bayesian method is not straightforward and is not included here. We also study the effects of different learning periods and different bias correction methods on the SLA performances. This paper is organized as follows. In Section II, we present the data that we used in this study, including the models and the reanalysis data. In addition, we discuss the different bias correction methods that we used. In Section III, we describe the SLAs and the regression forecasting methods as we implemented them. We also provide the details of the climatology that we derived from the reanalysis data. In Section V, we present the predictions of the different forecasting methods. We also evaluate their global and regional performances based on their root mean square errors (*RMSEs*). The global and regional uncertainties of the predictions of the different forecasting methods are presented in Section VI. The weights assigned by the SLAs to the different models and to the climatology (all the members of the ensemble) are presented in Section VII.



The results are discussed and summarized in Section VIII.

## II. MODELS AND DATA

The decadal experiments were introduced to the Coupled Model Intercomparison Projects (CMIP) multi-model ensemble in its fifth phase (CMIP5). The objective of these experiments is to investigate the ability of climate models to produce skillful future climate predictions for a decadal time scale. The climate models in these experiments were initialized with interpolated observation data of the ocean, sea ice and atmospheric conditions, together with the atmospheric composition [49]. The ability of these simulations to produce skillful predictions was not investigated widely, but it was shown that they can generate skillful predictions in specific regions around the world [4, 5, 10, 16, 51–55].

The CMIP5 decadal experiments were initialized every five years between 1961 and 2011 for 10-year simulations, with three exceptional experiments that were extended to 30-year simulations. One of these 30-year experiments was initialized in 1981 and simulated the climate dynamics till 2011. The output of four variables from this experiment is tested here—surface temperature, zonal and meridional surface wind components, and surface pressure. In what follows, we analyze the monthly means of these variables.

Table I shows the eight climate models included in our ensemble. The decadal experiments of the CMIP5 project include a set of runs for each of the models, differing by the starting date and the initialization scheme used. We chose, arbitrarily, the first run of each model. As long as the model variability is the main source of uncertainty, the choice of the realization should not be significant for our analysis. Indeed, it was found that, in the CMIP5 decadal experiments, the model variability is the main source of uncertainty, independent of the prediction lead time, as long as the predictions are not bias corrected. Bias correction reduces mainly the model variability; however, the contribution of the model variability remains important [18].

The NCEP/NCAR reanalysis data [50] were used as the observation data for the learning and for the evaluation of the forecasting methods performances. We are aware of other reanalysis projects [56, 57]; however, we selected the NCEP based on its wide use (note that the assessment of the quality of the different reanalysis projects is subjective and is beyond the scope of this paper). The effects of using different reanalysis data are left for future research.

TABLE I: Model Availability

Institute ID	Model Name	Modeling Center (or Group)	Grid (lat X lon)
BCC	BCC-CSM1.1	Beijing Climate Center, China Meteorological Administration	64 X 128
CCCma	CanCM4	Canadian Centre for Climate Modelling and Analysis	64 X 128
CNRM-CERFACS	CNRM-CM5	Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	128 X 256
LASG-IAP	FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	108 X 128
IPSL*	IPSL-CM5A-LR	Institute Pierre-Simon Laplace	96 X 96
MIROC	MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	128 X 256
	MIROC4h		320 X 640
MRI	MRI-CGCM3	Meteorological Research Institute	160 X 320

\* not available for U and V components of wind

#### A. Bias correction

The predictions made by the climate models often suffer from inherent systemic errors [58], and it is common to apply bias correction methods to the model outputs before analyzing them. For long-term climate projections, this procedure is more straightforward because of the available reference period. Bias correction in decadal climate predictions is not trivial not only because there is no clear reference period but also because some of these experiments are known to have a drift from the initial condition to the model's climatology during the first years of the simulation [4].

Here, two bias correction methods and the original data were considered. The original data without any bias correction is noted as *no correction*. The first bias correction method corresponds to subtracting from each model results their average during the learning period and adding the climatological average (the average of the NCEP/NCAR reanalysis data for the same period). This method is noted as *average correction*. The second bias correction method corresponds to subtracting from each model and for each calendar month the corresponding average during the learning period and adding the NCEP/NCAR reanalysis average for that calendar month during the same learning period. This method is noted as *climatology correction*. The two bias correction methods described above do not account for the explicit time dependence of the bias. However, it is reasonable to assume that for decadal climate predictions, the bias does not change considerably with time.

### III. FORECASTING METHODS

In this work, we consider three sequential learning algorithms (SLAs), introduced below. More thorough descriptions of the SLAs can be found in Ref. [43] and in Ref. [59]. We also consider the linear regression (REG) [37] method in order to compare the performances of the SLAs to the well-known regression method. The climatology (CLM) is considered here as the threshold for skillful predictions. For clarity, the equations that describe the forecasting methods omit the spatial indices. However, the forecasting schemes were applied to each of the grid cells independently, thereby allowing the spatial distribution of the weights (or the coefficients in the case of the REG) and the reference climatology.

#### A. The EWA and the EGA

The SLAs use an ensemble of *experts* (climate models), each of which provides a prediction for a future value of a climate variable, to provide a forecast of the climate variable in terms of the weighted average of the ensemble. The process is sequentially repeated with the weights of the models being updated, after each measurement, according to their prediction skill. We divide the period of the model simulations into two parts. The first part is the learning (or training) period whose data is used to update the model weights in the manner described above, and the second part is used for validating and evaluating the *forecaster* performance. At the end of the learning

period, the learning ends and the weights generated by the SLA in the last learning step are used to weight the predictions of the climate models during the validation period.

The deviation of the prediction of model  $E$ ,  $f_{E,t}$ , from the observed value,  $y_t$ , determines the *loss function*,  $l(f_{E,t}, y_t)$ , at time  $t$ . Similarly, the loss function of the *forecaster* (the SLA) is determined by the deviation of its prediction,  $p_t$ , from the observed value at time  $t$ . The *loss function* is the metric used to evaluate the models performances. In our study, we define the *loss function* as the square of the deviation, namely,  $l(f_{E,t}, y_t) \equiv (f_{E,t} - y_t)^2$  for model  $E$  and  $l(p_t, y_t) \equiv (p_t - y_t)^2$  for the *forecaster*.

The output of the Exponentiated Weighted Average (EWA), the first SLA described here, at time  $t$  is the set of the weights of the models in the ensemble:

$$w_{E,t}^{EWA} \equiv \frac{1}{Z_t} \cdot w_{E,t-1}^{EWA} \cdot e^{-\eta \cdot l_{E,t}} \quad (1)$$

where  $\eta$  is a positive number representing the learning rate of the *forecaster* and  $Z_t$  is a normalization factor. The EWA prediction at time  $t$  is defined below:

$$p_t^{EWA} \equiv \sum_{E=1}^{N_e} w_{E,t-1}^{EWA} \cdot f_{E,t}, \quad (2)$$

where  $N_e$  is the number of models in the ensemble.

The second SLA considered here is the Exponentiated Gradient Average (EGA). The EGA assigns the weights according to the following rules:

$$w_{E,t}^{EGA} \equiv \frac{1}{Z_t} \cdot w_{E,t-1}^{EGA} \cdot e^{-\eta \cdot l'_{E,t}}, \quad (3)$$

where  $l'_{E,t}$  is the gradient of the *forecaster loss function* with respect to the weight of model  $E$  at time  $t - 1$ . The mathematical definition of  $l'_{E,t}$  is provided below:

$$l'(f_{E,t}, p_t^{EGA}, y_t) \equiv \frac{\partial l(p_t^{EGA}, y_t)}{\partial w_{E,t-1}^{EGA}} = 2 \cdot (p_t^{EGA} - y_t) \cdot f_{E,t}, \quad (4)$$

where the prediction of the EGA,  $p_t^{EGA}$ , is defined similarly to the prediction of the EWA:

$$p_t^{EGA} \equiv \sum_{E=1}^{N_e} w_{E,t-1}^{EGA} \cdot f_{E,t}. \quad (5)$$

An important difference between the EWA and the EGA is the fact that under ideal conditions and stationary time series, the EWA converges to the best model in the ensemble, while the EGA converges to the observations [48].

Note that for the first learning step, one has to assign initial weights to the models. Without any a priori knowledge of the models performances, the natural choice is to assign equal weights to all the models. If the hierarchy of the models is known, it is possible to assign their initial weights accordingly.

The learning rate,  $\eta$ , was optimized by scanning a wide range of values and using the value that resulted in the minimal  $RMSE$  during the learning period. However, we added a restriction that the maximal change in the weight of each of the models, between two learning steps, will be smaller than the weight of each model in an equally weighted ensemble—namely,  $1/N_e$ . This restriction was added to ensure the stability of the weights. The metric that we used for this optimization is defined below:

$$M \equiv RMSE \cdot \left( 1 + \Theta \left( \max_{E=1,\dots,N_e, t=1,\dots,n} \frac{\Delta w_{E,t}}{(1/N_e)} - 1 \right) \right), \quad (6)$$

where  $\Theta$  represents the Heaviside theta function, and  $RMSE$  is the root mean squared error of the *forecaster* during the  $n$  time steps of the learning period. The  $RMSE$  for a grid cell  $(i, j)$  is conventionally defined.

$$RMSE(i, j) \equiv \sqrt{(1/n) \sum_{t=1}^n (p_t(i, j) - y_t(i, j))^2}. \quad (7)$$

The value of  $\eta$  that minimizes  $M$  was found using a recursive search within a very wide range of values restricted only by the machine precision. The optimization was done for each grid cell separately.

## B. The Learn- $\alpha$ algorithm

The basic form of the EWA was modified to explicitly allow switching between *experts*. This switching improves the performance of the SLA when dealing with nonstationary time series. The fixed-shared algorithm introduced in Ref. [60] is defined by the following rules:

$$w_{E,t+1}^{FSA} = \frac{1}{Z_t} \cdot \sum_{E^*=1}^{N_e} w_{E,t}^{FSA} \cdot e^{-\eta \cdot l_{E^*,n}} \cdot K(E, E^*), \quad (8)$$

where

$$K(E, E^*; \alpha) \equiv (1 - \alpha) \cdot \delta(E, E^*) + \frac{\alpha}{N_e - 1} \cdot (1 - \delta(E, E^*)). \quad (9)$$

Here,  $\alpha \in [0, 1]$  is the switching rate parameter, and  $\delta(\cdot, \cdot)$  is the Kronecker delta.

The fixed-share algorithm was extended in Ref. [59] by also learning the optimal switching rate parameter,  $\alpha$ . This modified SLA is known as the Learn- $\alpha$  algorithm (LAA). In the LAA, the algorithm scans a range of switching rates,  $\alpha_j$ ,  $j \in 1, \dots, N_\alpha$ , and assigns weights to each value of  $\alpha_j$  based on a loss per alpha function,  $l_t(\alpha_j) \equiv -\log\left(\sum_{E=1}^{N_e} w_{E,t}(\alpha_j) e^{-l_{E,t}}\right)$ . The weights are updated sequentially for both the switching rate and the *experts*. The updating rule for the weight of a specific value,  $\alpha_j$ , is provided below:

$$W_t(\alpha_j) = \frac{1}{Z_t} W_{t-1}(\alpha_j) e^{-l_t(\alpha_j)}. \quad (10)$$

The updating rule for the weight of *expert*  $E$ , given  $\alpha_j$ , is provided below:

$$w_{E,t}^{LAA}(\alpha_j) = \frac{1}{Z_t(\alpha_j)} \sum_{E^*=1}^{N_e} w_{E^*,t-1}^{LAA}(\alpha_j) e^{-l_{E^*,t} K(E, E^*; \alpha_j)}. \quad (11)$$

The prediction at time  $t$ , is the weighted average of the *experts* and the different values of  $\alpha$ .

$$p_t^{LAA} = \sum_{E=1}^{N_e} \sum_{j=1}^{N_\alpha} W_{t-1}(\alpha_j) \cdot w_{E,t-1}^{LAA}(\alpha_j) \cdot f_{E,t}. \quad (12)$$

Here, we adopted a discretization of  $\alpha$  to optimize the LAA performance [59].

### C. Regression

The linear regression algorithm considered here is described in Ref. [37]. In this algorithm, the forecast is a linear combination of the climate model predictions as described below:

$$p_t^{REG} = \bar{y} + \sum_{E=1}^N a_E (f_{E,t} - \bar{f}_E). \quad (13)$$

Here,  $\bar{y} \equiv (1/n) \sum_{t=1}^n y_t$  is the temporal mean of the observed values during the learning period (similarly,  $\bar{f}_E$  is the temporal mean value of the predicted values by *expert*  $E$  during the learning period), and  $a_E$  are the regression coefficients minimizing the sum of squared errors during the learning period,  $G$ , which is defined below:

$$G \equiv \sum_{t=1}^n (p_t - y_t)^2, \quad (14)$$

where  $n$  is the number of time steps in the learning period. The algorithm that we used to minimize  $G$  involved the elimination of models that were linearly dependent on the other models in the ensemble.

#### D. Climatology

The climatology is defined here as the monthly averages of the observed conditions during the learning period. Namely,

$$C_m = \sum_{t=1}^{n_1} y_{t,m} \quad (15)$$

where  $y_{t,m}$  is the observed value in month  $m \in [1, 12]$  of year  $t$  ( $t$  is measured in years from the beginning of the simulations), and  $n_1$  is the duration of the learning period in years (for simplicity, we assume here that both the learning and the validation periods span an integer number of years). The twelve months of the climatology were replicated to match the duration of the validation period; that is,

$$CLM_{t,m} = C_m, \quad (16)$$

for  $t \in [n_1 + 1, n_1 + n_2]$  ( $n_2$  is the duration of the validation period in years). The climatology is often considered as the threshold for a skillful prediction, i.e., a *forecaster* that outperforms the climatology is considered skillful.

#### IV. EVALUATION METRICS

Two main evaluation metrics are used here: the average error, quantified by the *RMSE* of each of the *forecasters*, and the variability of the ensemble predictions, characterized by their standard deviation, the *STD*. The global averages of the *RMSE* and the *STD* are calculated by weighting each grid cell by the fraction of the earth's surface it spans. The precise details are provided here for clarity. During the validation period, the *RMSE* of each *forecaster* was calculated for each grid cell (because all the climate variables studied here are two-dimensional, each grid cell has two indices,  $(i, j)$ ) from the time series of the forecast and the observations. Then, the global area-weighted average of the *RMSE* ( $RMSE_{GAW}$ ) was calculated as detailed below:

$$RMSE_{GAW} \equiv (1/A_{Earth}) \sum_{i,j} A_{i,j} RMSE(i, j), \quad (17)$$

where  $A_{Earth}$  is the total earth's surface area, and  $A(i, j)$  is the area spanned by the  $(i, j)$  grid cell. In what follows, we will present both the spatial distribution of the *RMSE* and its global average.

Similarly to the *RMSE*, the variance of the ensemble predictions was calculated for each of the grid cells at each time point and then averaged over time during the validation period. The square

root of this temporally averaged variance is what we define here as the *STD* of each grid cell. The mathematical definition of the *STD* is provided below:

$$STD(i, j) \equiv \sqrt{(1/n) \sum_{t=1}^n \sum_{E=1}^N w_E(i, j) (f_{E,t}(i, j) - p_t(i, j))^2}. \quad (18)$$

The global area-weighted average was then calculated:

$$STD_{GAW} \equiv (1/A_{Earth}) \sum_{i,j} A_{i,j} STD(i, j). \quad (19)$$

The skill of the *forecasters* was measured by comparing their *RMSE* and *STD* to those of some other reference *forecaster*. For convenience, we define below the *RMSE* skill score,  $R_{ref, fct}$ :

$$R_{ref, fct} \equiv \frac{RMSE_{ref} - RMSE_{fct}}{\frac{1}{2} (RMSE_{ref} + RMSE_{fct})}. \quad (20)$$

The indices *ref* and *fct* are used to identify the *forecasters* whose skills are compared. Similarly, we define below the *STD* skill score,  $S_{ref, fct}$ :

$$S_{ref, fct} \equiv \frac{STD_{ref} - STD_{fct}}{\frac{1}{2} (STD_{ref} + STD_{fct})}. \quad (21)$$

Unless otherwise specified, we used the climatology as the reference *forecaster* for  $R_{ref, fct}$  and the equally weighted ensemble as the reference *forecaster* for  $S_{ref, fct}$ . Note that the skill scores are defined such that a *forecaster* with a smaller *RMSE* than the reference *forecaster* has a positive  $R_{ref, fct}$  score, and similarly, a *forecaster* with a smaller *STD* (i.e., smaller uncertainty) than the reference *forecaster* has a positive  $S_{ref, fct}$  score.

## V. PREDICTIONS

### A. Global

The simplest measure of the performance of the *forecasters* is the global average of the root mean squared error,  $RMSE_{GAW}$ . Figure 1 shows the  $RMSE_{GAW}$  of the validation period for the five different *forecasters*, EWA, EGA, REG, LAA and CLM, and the different learning periods. The rows (from top to bottom) correspond to the surface temperature, zonal wind, meridional wind, and pressure, respectively. The columns (from left to right) correspond to no bias correction, average bias correction, and climatology bias correction, respectively. The data is provided in



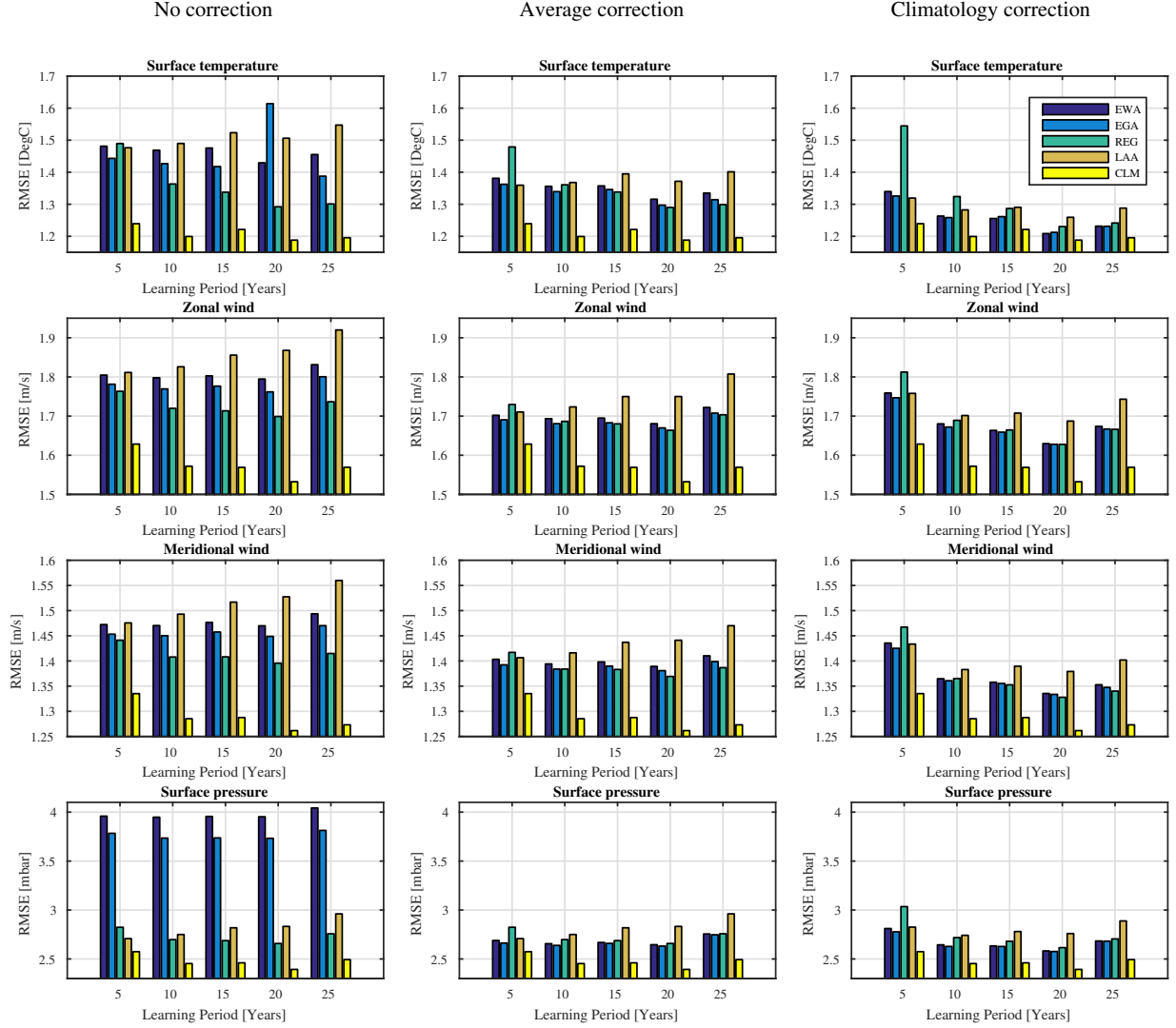


FIG. 1: Globally averaged  $RMSE$ .  $RMSE_{GAW}$  for the five forecasting methods (EWA, EGA, REG, LAA and CLM), learning periods of 5, 10, 15, 20 and 25 years and the four climate variables (surface temperature, two wind components and pressure). The ensemble used by the *forecasters* does not include the climatology. The left panels correspond to no bias correction, the middle panels correspond to average bias correction, and the right panels correspond to climatology bias correction (see Section II A for the details of the different bias correction methods).

Tables 1-4 of the Supplementary Information. The decadal climate simulations considered here span a 30-year period that is split such that the first part is used for learning and the second part is used for the evaluation of the performances; that is, for the five-year learning period, the validation period is the next 25 years, and for the 10-year learning period, the validation period is the next

20 years, etc. The  $RMSE_{GAW}$ s of the individual models are not presented because they are much higher than those of the *forecasters*. The  $RMSE_{GAW}$  of the equally weighted ensemble is much lower than those of the models, but it is also too high to be included within the scale shown in Fig. 1. The bias correction that resulted in the smallest  $RMSE_{GAW}$ s is the *climatology correction*, which is described in Section II A.

Figure 1 shows that the climatology outperforms all the other *forecasters*, for all the learning periods and bias correction methods studied here. Therefore, we added the climatology as an *expert* to the ensemble. Unless otherwise specified, the following results were derived from an ensemble including the climatology as an additional *expert* [48].

Figure 2 shows the same results as Figure 1 for an ensemble that includes the climatology. In addition, the initial weight assigned to the climatology was 0.5, whereas the initial weight of all the other models was  $0.5/(N_e - 1)$  ( $N_e - 1$  is the number of the models excluding the climatology). This higher initial weight of the climatology was motivated by its superior performance (as shown in Fig. 1 and [48]). The data that was used to generate Fig. 2 is provided in Tables 5-8 of the Supplementary Information.

The results of Fig. 2 show that the best predictions are obtained using 20 years of learning and different bias correction methods for different variables and different *forecasters*. The fact that the  $RMSE_{GAW}$  is minimized after 20 years of learning can be related to two factors: i) for short learning periods, there is a longer prediction period and, therefore, a larger  $RMSE_{GAW}$ ; ii) for the 25-year learning period, the time lead from the initialization to the prediction period is long, and in addition, the short five-year prediction period does not represent the climate variability over a time scale of 25 years (the duration of the learning period). The 20 years of learning also ensures that the learning period extends well beyond the drift of the models. In Table II, we detail the bias correction that resulted in the smallest  $RMSE_{GAW}$  for each *forecaster* and for each climate variable. In what follows, we will present only the results of these bias corrections and 20 years of learning. We find that all the SLAs have a lower or equal  $RMSE_{GAW}$  than the climatology for the surface temperature and wind components. For the surface pressure, only the LAA outperforms the climatology. We also see that, for most climate variables, the  $RMSE_{GAW}$ s of the EWA and the climatology are almost equal. This is not a coincidence; it reflects the fact that the EWA tracks the best model, which in most grid cells, is the climatology. The two other SLAs reduce the  $RMSE_{GAW}$  below that of the climatology by extracting information from the other models in the ensemble. The LAA outperforms the EGA for short learning periods ( $< 15$  years) and

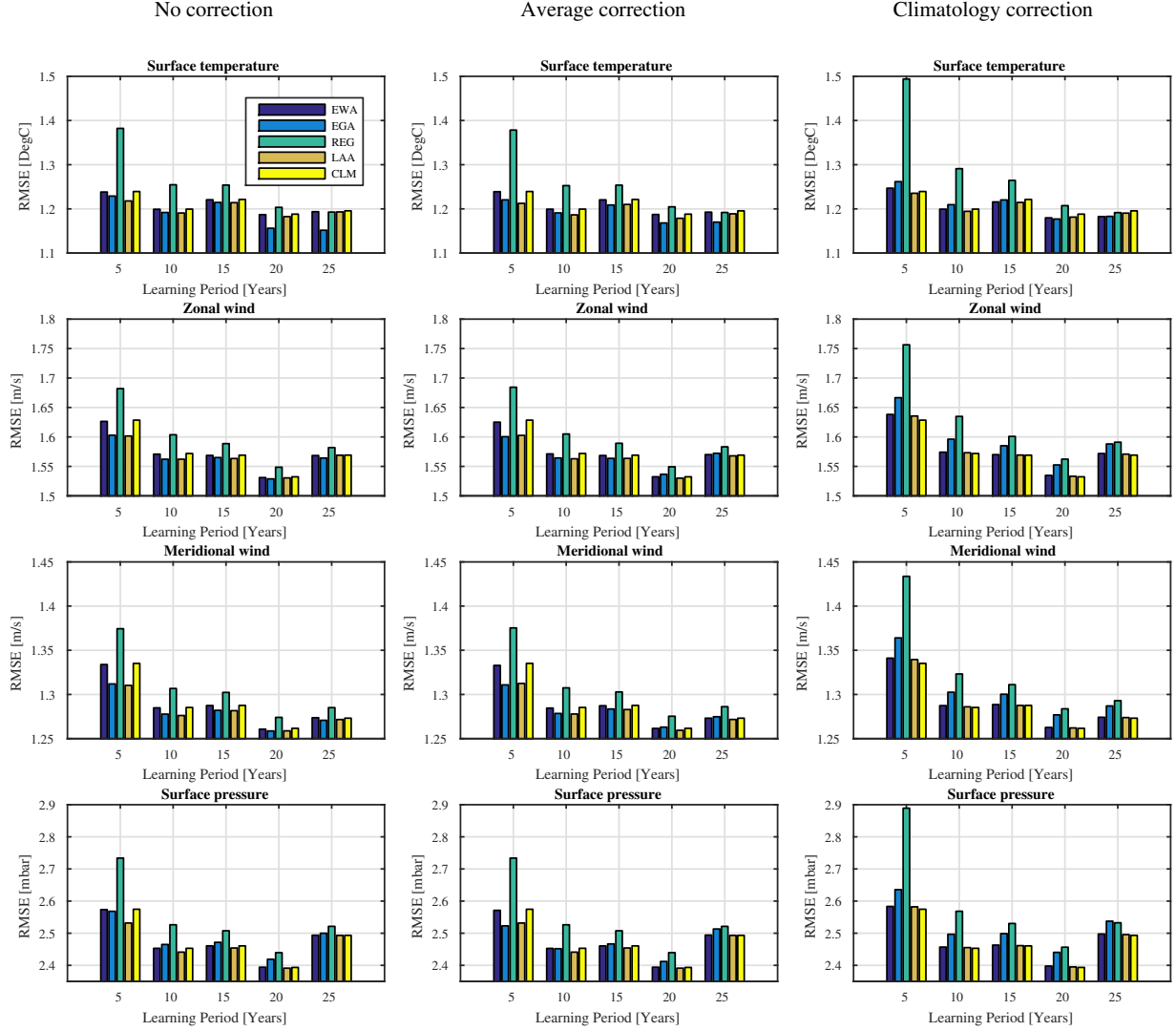


FIG. 2: Globally averaged  $RMSE$  with climatology.  $RMSE_{GAW}$  for the five forecasting methods (EWA, EGA, REG, LAA and CLM), learning periods of 5, 10, 15, 20 and 25 years and the four climate variables (surface temperature, two wind components and pressure). The ensemble used by the *forecasters* includes the climatology. The left panels correspond to no bias correction, the middle panels correspond to average bias correction, and the right panels correspond to climatology bias correction.

for all learning periods in the predictions of the surface pressure. This better performance can be attributed to the design of the LAA for the learning of nonstationary data. The poorer performance, relative to the climatology, of most of the *forecasters* (except for the LAA) in the prediction of the surface pressure is not fully understood. However, we found that for the surface pressure, the variability between the models is often larger than its seasonal variability, while all the other

climate variables considered here show seasonal variabilities that are larger than the variabilities between the models. It is also possible that the model predictions of the monthly mean surface pressure are worse than the predictions of the other climate variables.

TABLE II: The optimal bias correction for each *forecaster* and each climate variable.  $T$ ,  $U$ ,  $V$ , and  $P$  denote the surface temperature, zonal wind, meridional wind and pressure, respectively. *nbias*, *bias* and *mbias* correspond to the *no correction*, *average correction* and *climatology correction*, respectively.

Forecaster	Climate variable			
	T	U	V	P
EGA	<i>nbias</i>	<i>nbias</i>	<i>nbias</i>	<i>bias</i>
EWA	<i>mbias</i>	<i>nbias</i>	<i>nbias</i>	<i>nbias</i>
LAA	<i>bias</i>	<i>bias</i>	<i>nbias</i>	<i>bias</i>
REG	<i>nbias</i>	<i>nbias</i>	<i>nbias</i>	<i>nbias</i>
AVG	<i>mbias</i>	<i>mbias</i>	<i>mbias</i>	<i>mbias</i>

## B. Regional

The  $RMSE_{GAW}$  is convenient because it aims to quantify the performances of the *forecasters* using only one number. However, often the more scientifically and practically relevant information are the spatial distributions of the  $RMSE$ . In this subsection, the spatial distribution of the *forecaster* performances will be investigated using the  $R_{ref, fct}$  metric defined above. This metric will allow us to compare the performances of the different *forecasters* and, in particular, to compare their performances to that of the trivial *forecaster*—the climatology. The statistical significance of the improvement achieved by the *forecasters* was tested by introducing the null hypothesis that the temporal distribution of  $R_{ref, fct}$  is symmetric around 0. Grid cells in which the hypothesis was rejected with a 90% confidence level in favor of a better *forecaster* performance are marked with white dots. Similarly, grid cells in which the hypothesis was rejected in favor of a poorer *forecaster* performance are marked with black dots. Grid cells in which the data does not provide enough evidence to reject the null hypothesis are not marked.

Figure 3 depicts the spatial distributions of  $R_{CLM, EGA}$  (upper left panel),  $R_{CLM, EWA}$  (upper

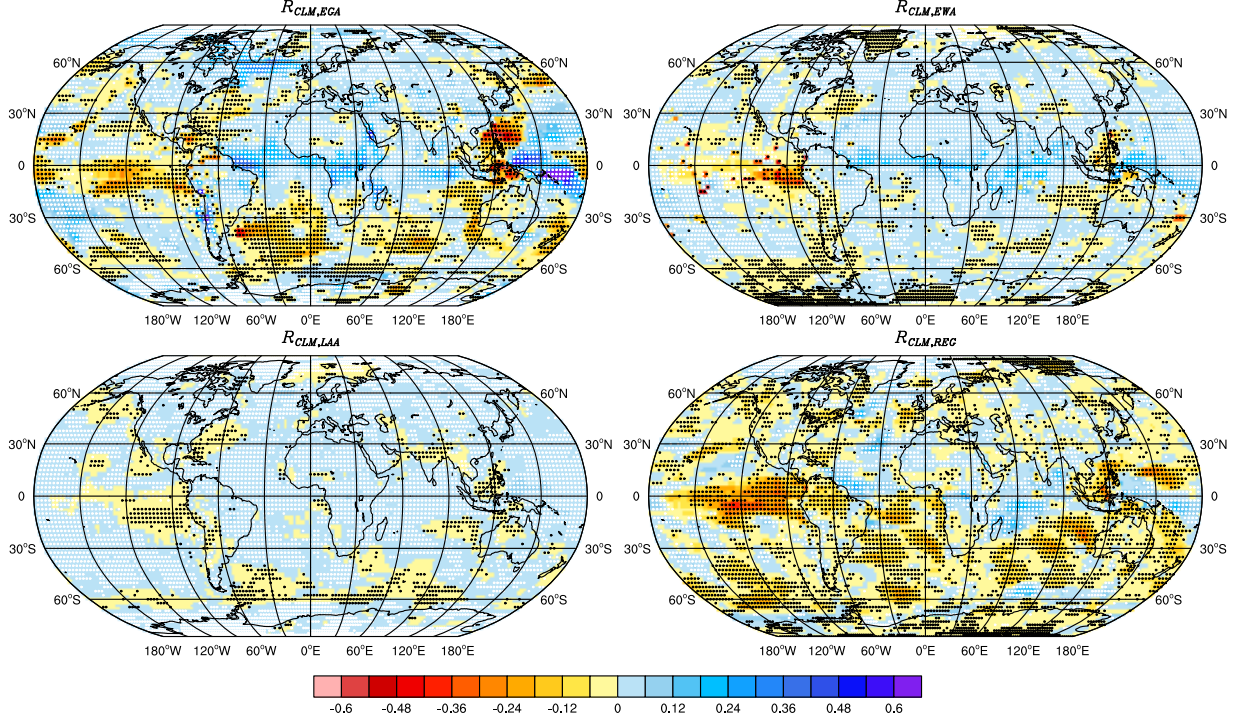


FIG. 3: Surface temperature  $RMSE$  skill score. Upper left panel: EGA, upper right panel: EWA, lower left panel: LAA and lower right panel: REG. Positive values correspond to a smaller  $RMSE$  than the climatology and vice versa. White circles represent significant improvement and black circles represent a significantly poorer performance.

right panel),  $R_{CLM,LAA}$  (lower left panel) and  $R_{CLM,REG}$  (lower right panel) for the surface temperature. This figure better clarifies the origin of the EGAs superior performance over the other *forecasters* (as seen from the surface temperature panels, the 20-year learning period bins of Fig. 2). The largest variability is observed for  $R_{CLM,EGA}$  and the smallest variability for  $R_{CLM,LAA}$ . While the LAA shows a positive skill score over large regions, the score is relatively low, reflecting a small improvement in the prediction compared with the climatology. For the EGA, on the other hand, we see that over regions in the North Atlantic, South America, central Africa, and Oceania, there is a large improvement relative to the climatology, while in regions in the East China Sea, the South Atlantic Ocean and the Eastern Central Pacific Ocean, there is a much poorer performance compared with the climatology. The regression *forecaster* shows a poorer performance compared with the climatology (negative skill score) over most of the globe. All the *forecasters* show a positive skill over regions in North Africa, Asia and North America, suggesting that the models are capable of capturing deviations from the climatology in these regions.

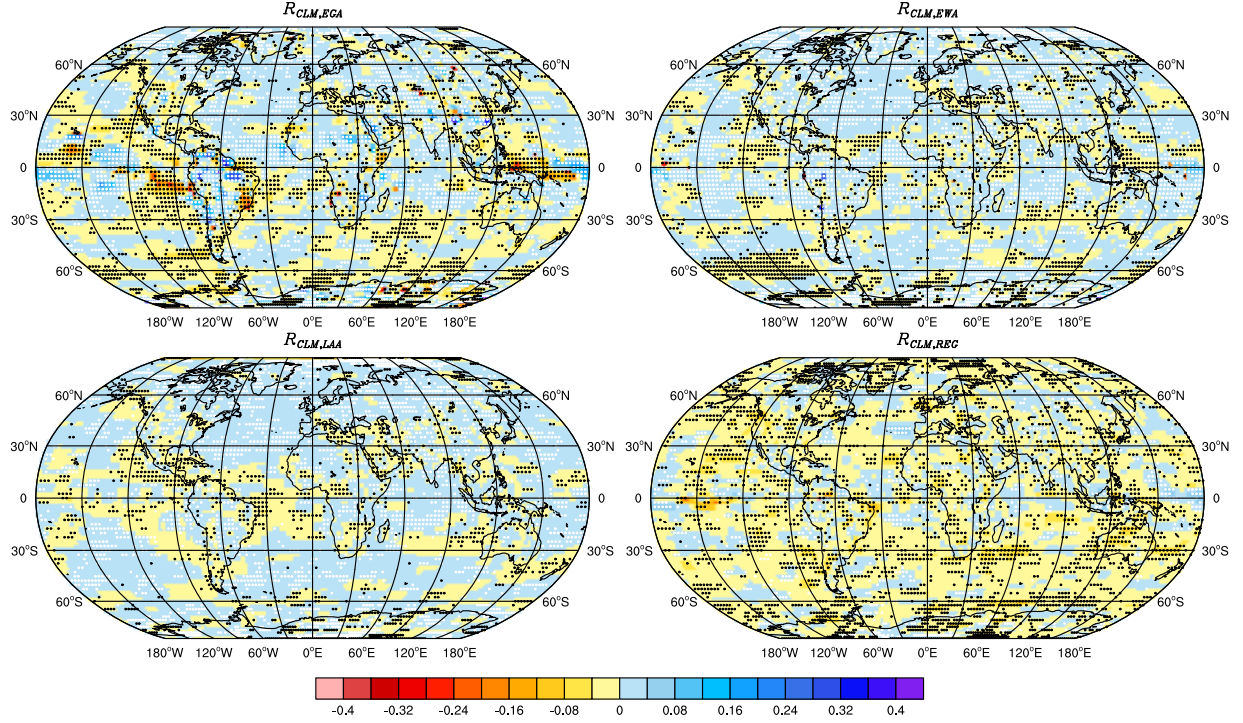


FIG. 4: Surface zonal wind  $RMSE$  skill score. Upper left panel: EGA, upper right panel: EWA, lower left panel: LAA and lower right panel: REG. Positive values correspond to a smaller  $RMSE$  than the climatology and vice versa. White circles represent significant improvement and black circles represent a significantly poorer performance.

The spatial distribution of the  $RMSE$  skill score for the zonal and meridional wind components are shown in Figures 4 and 5, respectively. Both wind components have similar characteristics. The EGA shows a similar distribution of the skill for the wind components to that found for the surface temperature. The EWA and the LAA show almost zero skill over most of the globe due to the fact that they both assign a very high weight to the climatology and a very small weight to the other models. Although the improvement relative to the climatology is small, it was found to be statistically significant in many regions. The REG shows a poorer performance compared with the climatology over most of the globe.

Figure 6 shows the spatial distribution of the surface pressure  $R_{CLM,EGA}$  (upper left panel),  $R_{CLM,EWA}$  (upper right panel),  $R_{CLM,LAA}$  (lower left panel) and  $R_{CLM,REG}$  (lower right panel). The EGAs performance for the surface pressure is poor compared with its performance for the other variables. Large regions in the Pacific and Indian Oceans show a larger  $RMSE$  of the EGA than the climatology, while in some regions in the Atlantic Ocean, North Euro-Asia, Greenland



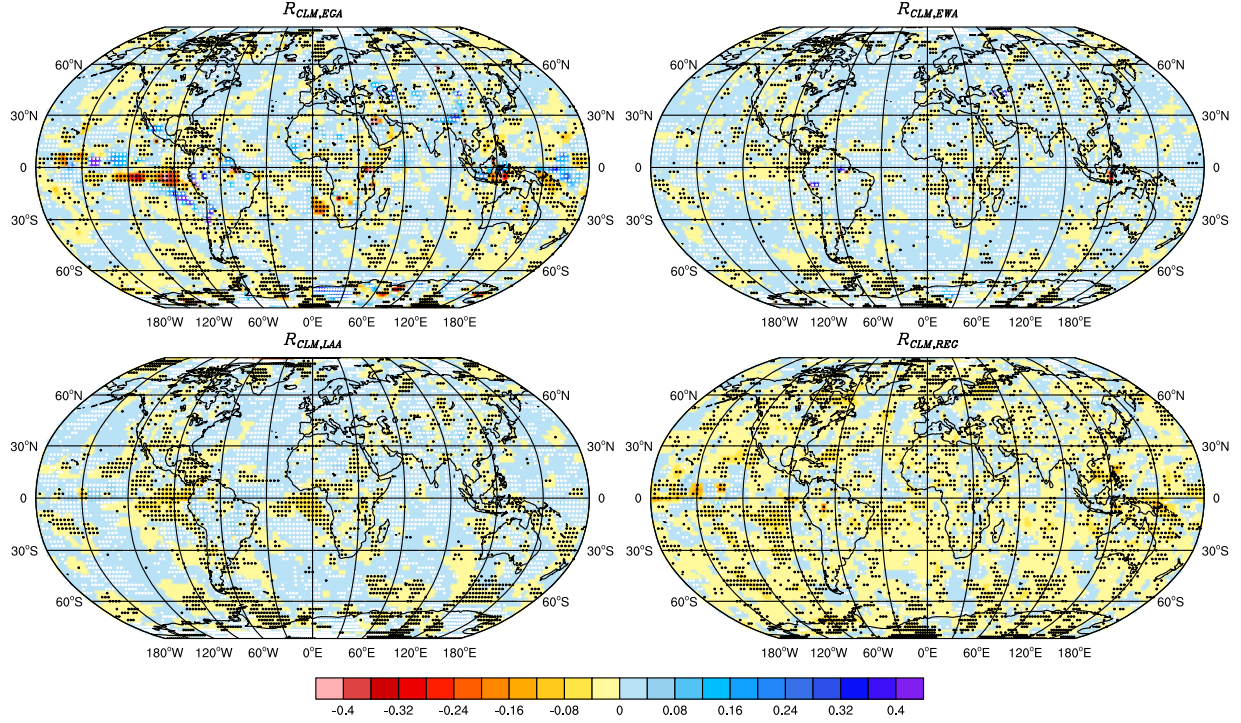


FIG. 5: Surface meridional wind  $RMSE$  skill score. Upper left panel: EGA, upper right panel: EWA, lower left panel: LAA and lower right panel: REG. Positive values correspond to a smaller  $RMSE$  than the climatology and vice versa. White circles represent significant improvement and black circles represent a significantly poorer performance.

and the South Pacific the EGA shows a better performance than the climatology. The EWA and LAA assign a very high weight to the climatology and, therefore, show an  $RMSE$  skill score close to zero. However, the small improvement achieved by the LAA is statistically significant over most of the globe. The REG shows a poorer performance than the climatology over most regions, with some exceptions in the central Atlantic Ocean and the Arabian Peninsula.

The EGA shows the highest  $RMSE$  skill score over most of the globe for the surface temperature and wind components, while the LAA shows the highest score for the surface pressure. There are several regions (such as the North Atlantic, North Indian Ocean and North Euro-Asia) where the SLAs seem to provide a smaller  $RMSE$  than the climatology. This suggests that at least some of the models capture processes that result in a deviation from the climatology and that the SLAs are capable of tracking these models.

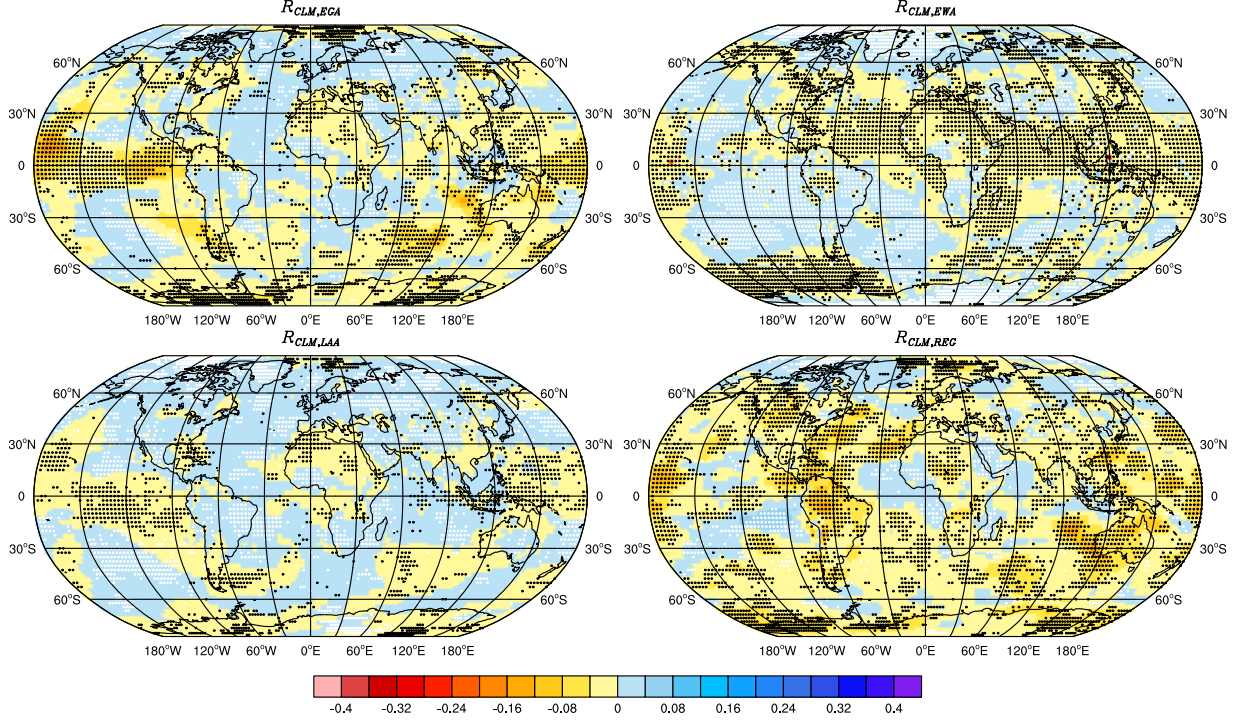


FIG. 6: Surface pressure  $RMSE$  skill score. Upper left panel: EGA, upper right panel: EWA, lower left panel: LAA and lower right panel: REG. Positive values correspond to a smaller  $RMSE$  than the climatology and vice versa. White circles represent significant improvement and black circles represent a significantly poorer performance.

## VI. UNCERTAINTIES

The  $RMSE$  is an important measure of the quality of the predictions; however, the uncertainties associated with the predictions of the *forecasters* are crucial for a meaningful assessment of the predictions quality. The uncertainties are quantified here using the standard deviation of the ensemble. A natural reference for comparing the variance of the ensemble weighted by the *forecasters* is the variance of the equally weighted ensemble that represents no learning. It was mentioned earlier that the linear regression does not assign weights to the models in the ensemble but rather attempts to find the linear combination of their predictions that minimizes the sum of squared errors. Therefore, in this section, we will compare the uncertainties of the three SLAs and the equally weighted ensemble, denoted here as AVR. Our analysis proceeds similarly to the analysis of the  $RMSE$ ; first we present the globally averaged standard deviation,  $STD_{GAW}$ , and then we present the spatial distribution of the  $STD$  skill score.



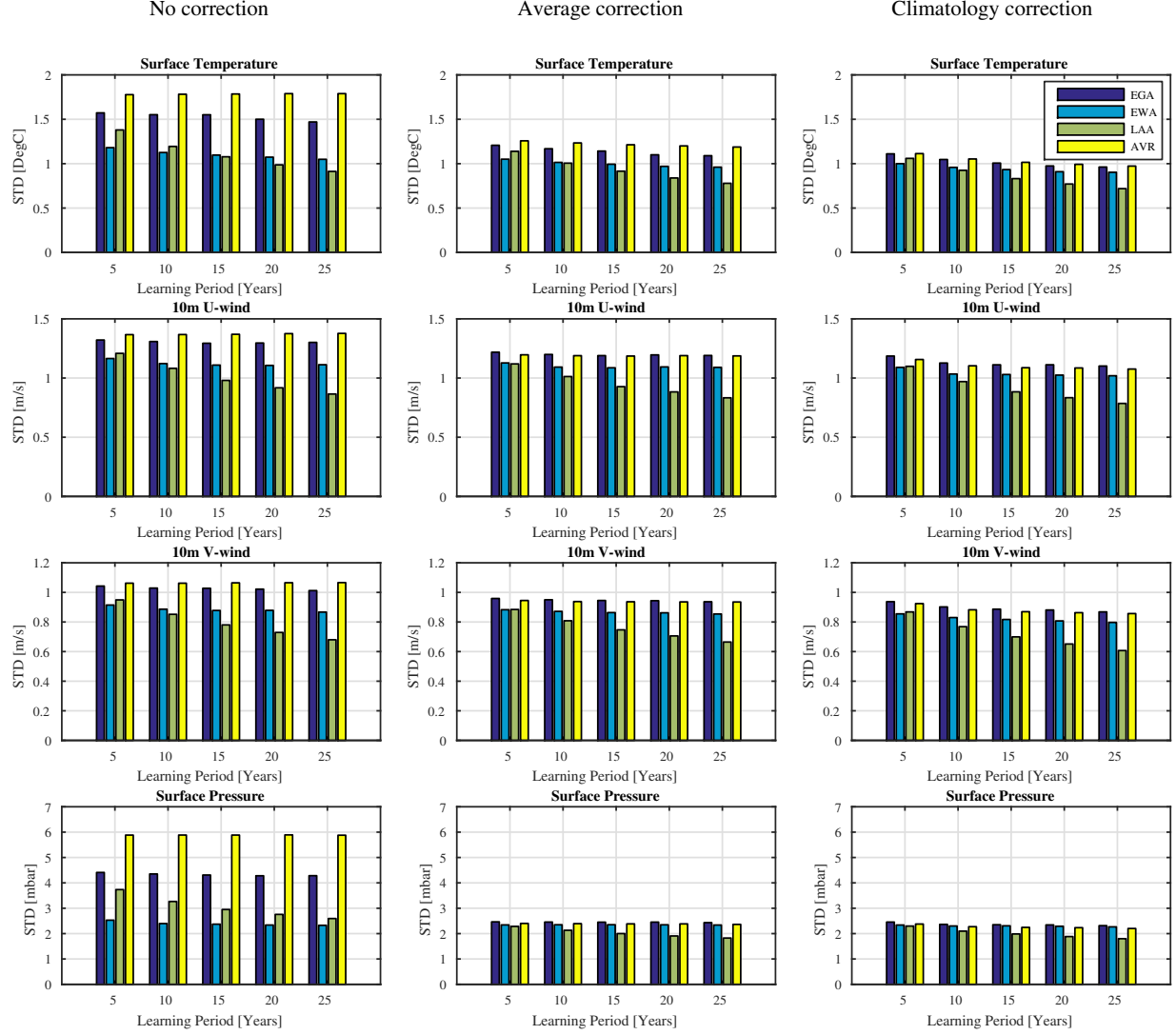


FIG. 7: Globally averaged  $STD$ .  $STD_{GAW}$  for the three SLAs (EWA, EGA, and LAA) and for the equally weighted ensemble, AVR, for learning periods of 5, 10, 15, 20 and 25 years and the four climate variables (surface temperature, two wind components and pressure). The ensemble used by the *forecasters* (and AVR) does not include the climatology. The left panels correspond to no bias correction, the middle panels correspond to average bias correction, and the right panels correspond to climatology bias correction (see Section II A for the details of the different bias correction methods).

### A. Global

Figure 7 shows  $STD_{GAW}$  of the EGA, EWA, LAA and AVR for different learning periods and for the four climate variables considered in this study. The results of Fig. 7 were derived from

an ensemble that does not include the climatology. The four left panels correspond to no bias correction, the four middle panels correspond to average bias correction and the four right panels correspond to climatology bias correction. The data is provided in Tables 9-12 of the Supplementary Information. As expected, the more detailed the bias correction, the smaller the uncertainty because it is associated with the anomaly rather than with the actual prediction. We also notice that without bias correction, the EWA has the smallest  $STD_{GAW}$ , while with bias correction, the LAA shows the smallest  $STD_{GAW}$ . Both the EWA and the LAA are expected to have lower  $STD$ s because they track the best models. With no bias correction, all the SLAs show smaller uncertainties than the equally weighted ensemble, while for the climatology bias correction, the EGA shows a higher  $STD_{GAW}$  than the AVR. This suggests that in large regions, the EGA assigns high weights to models spanning a broad range of predicted values. In addition, we notice that the  $STD_{GAW}$  is smaller for longer learning periods, or more precisely, for shorter prediction periods, as expected. The reduction of  $STD_{GAW}$  is more significant for the LAA because the longer learning allows it to better track the climatology despite the built-in switching rate.

Figure 8 is similar to Fig. 7 but for an ensemble that includes the climatology. The data used to generate Fig. 8 is provided in Tables 13-16 of the Supplementary Information. It is apparent that in this case, the  $STD_{GAW}$  of all the SLAs is smaller than that of the AVR, and for the longer learning periods, it is much smaller. The large reduction in the  $STD_{GAW}$  of the EWA and EGA is clearly associated with the fact that they track the climatology in most regions (because it is the best *expert* in these regions). The  $STD_{GAW}$  of the EGA is also reduced because it assigns a high weight to the climatology in many regions, but it still assigns significant weights to the other models; therefore, it has a larger  $STD_{GAW}$  than the other SLAs. The  $STD_{GAW}$  of the equally weighted ensemble does not change much because the climatology is only assigned a weight of  $1/N_e$ , and in most regions, the climatology is spanned by the other models.

## B. Regional

The uncertainty has a large spatial variability. We focus on the 20-year learning period and the ensemble that includes the climatology. The  $STD$  skill score shows the average temporal variability of the ensemble weighted by the *forecasters* compared with that of the equally weighted ensemble during the validation period. Figure 9 shows the spatial distribution of the surface temperature  $STD$  skill score for the three SLAs. The EGA has a positive  $STD$  skill score (smaller

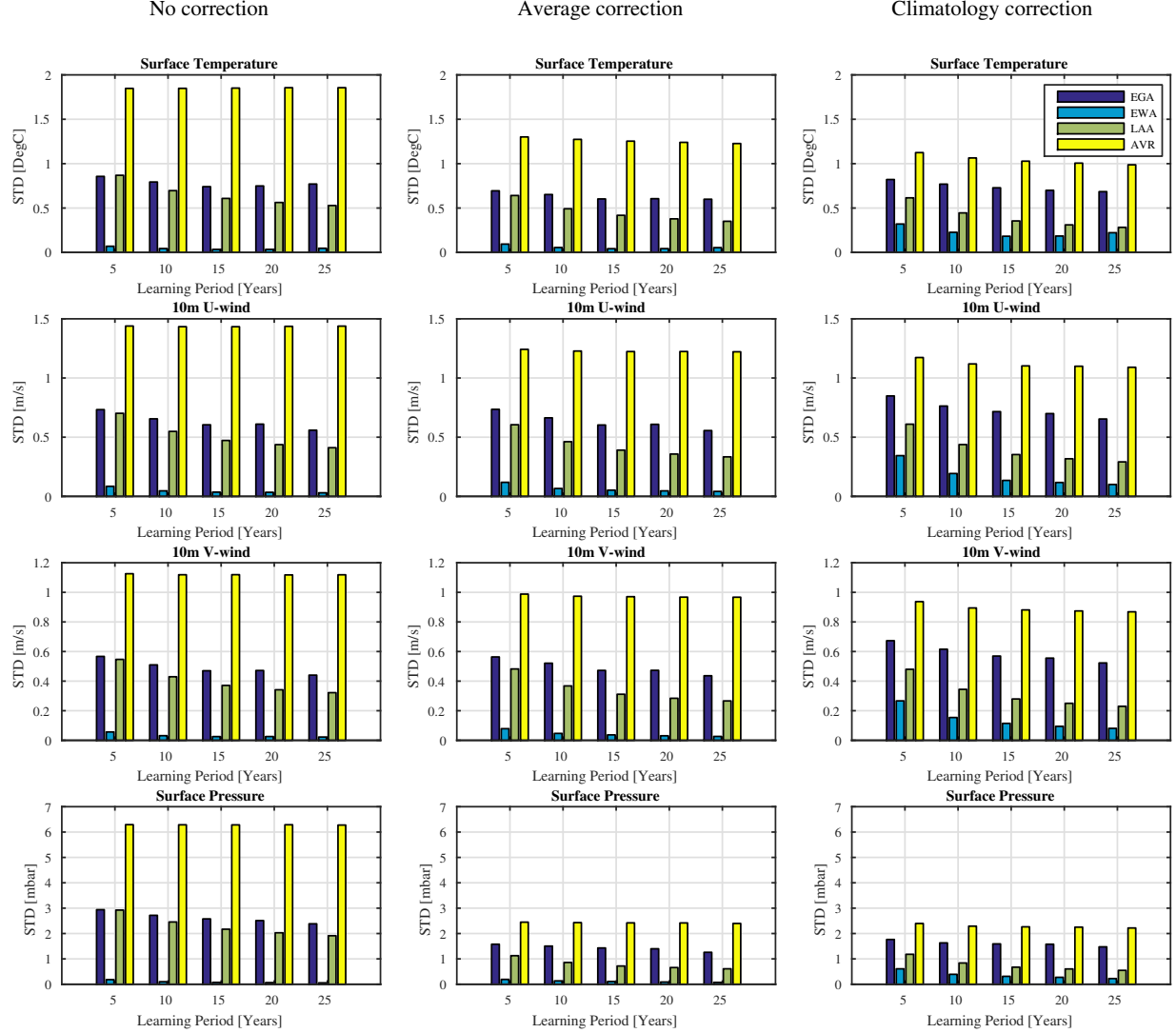


FIG. 8: Globally averaged  $STD$  with climatology.  $STD_{GAW}$  for the three SLAs (EWA, EGA, and LAA) and for the equally weighted ensemble, AVR, for learning periods of 5, 10, 15, 20 and 25 years and the four climate variables (surface temperature, two wind components and pressure). The ensemble used by the *forecasters* (and AVR) includes the climatology. The left panels correspond to no bias correction, the middle panels correspond to average bias correction, and the right panels correspond to climatology bias correction (see Section II A for the details of the different bias correction methods).

$STD$  than the equally weighted ensemble) in most of the globe, but there are many regions in which its  $STD$  is significantly larger than that of the AVR. The EWA reduces the  $STD$  over most of the globe except for the tropics. This reduction of the  $STD$  stems from the high weight assigned to the climatology. In many regions, the  $S_{AVR,EWA}$  is around 2, which reflects an almost

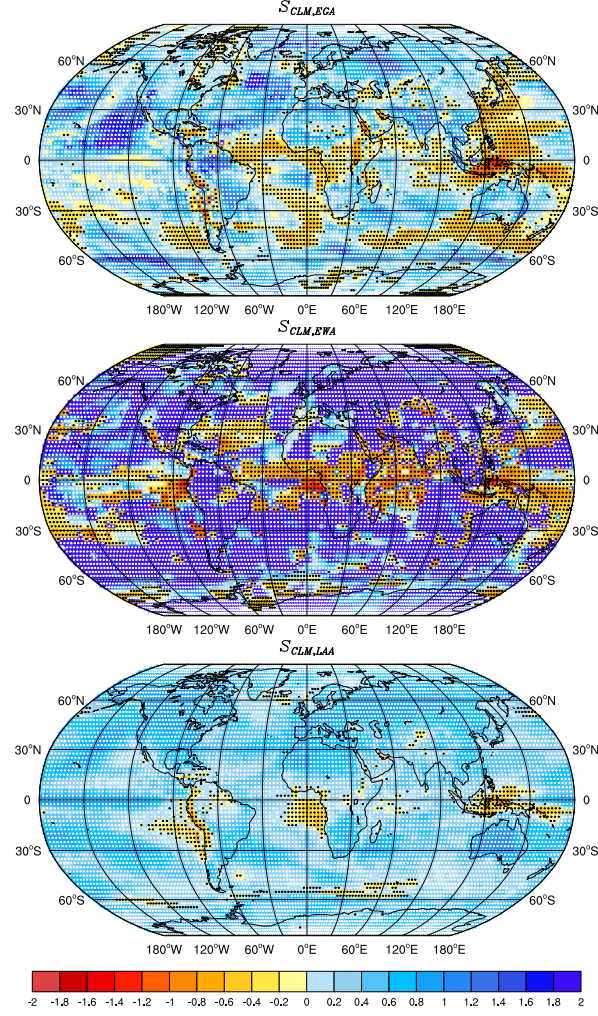


FIG. 9: Spatial distribution of the surface temperature  $STD$  skill score. Upper panel: EGA, middle panel: EWA, and lower panel: LAA. Positive values correspond to a smaller  $STD$  than the equally weighted ensemble and vice versa. White circles represent a statistically significant reduction of the  $STD$  and black circles represent a statistically significant increase of the  $STD$  relative to the  $STD$  of the equally weighted ensemble.

vanishing  $STD$  of the EWA. The LAA also shows a smaller  $STD$  than the AVR over most of the globe except for small regions in the tropics. Similarly to the EWA, the reduction of the uncertainties achieved by the LAA stems from the high weight assigned to the climatology. However, one can see that the  $S_{AVR,LAA}$  is smaller than the  $S_{AVR,EWA}$ , which reflects a lower weight of the climatology and higher weights of the other models due to the built-in switching rate in the LAA.

Figures 10 and 11 show the  $STD$  skill score of the EGA, EWA and LAA for the zonal and meridional wind components. For the wind components, all the SLAs show significant reductions

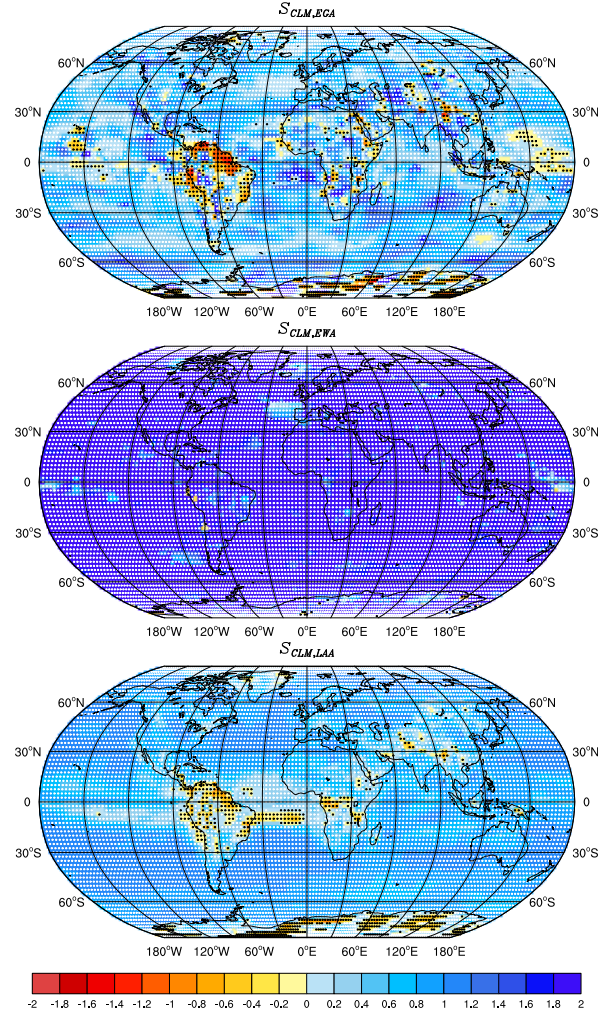


FIG. 10: Spatial distribution of the surface zonal wind  $STD$  skill score. Upper panel: EGA, middle panel: EWA, and lower panel: LAA. Positive values correspond to a smaller  $STD$  than the equally weighted ensemble and vice versa. White circles represent a statistically significant reduction of the  $STD$  and black circles represent a statistically significant increase of the  $STD$  relative to the  $STD$  of the equally weighted ensemble.

of the  $STD$  over most of the globe. The EGA and LAA show larger  $STD$ s in some small regions in the tropics. The results suggest that all the SLAs assign a high weight to the climatology, with the EWA almost fully converging to it, while the EGA and LAA extract information from the models as well.

Figure 12 shows the surface pressure  $STD$  skill score for the three SLAs. The EWA and LAA show positive skill scores over the entire globe, and the EGA only shows negative skill in a very small region in Oceania. The EWA fully converges to the climatology and has a vanishing

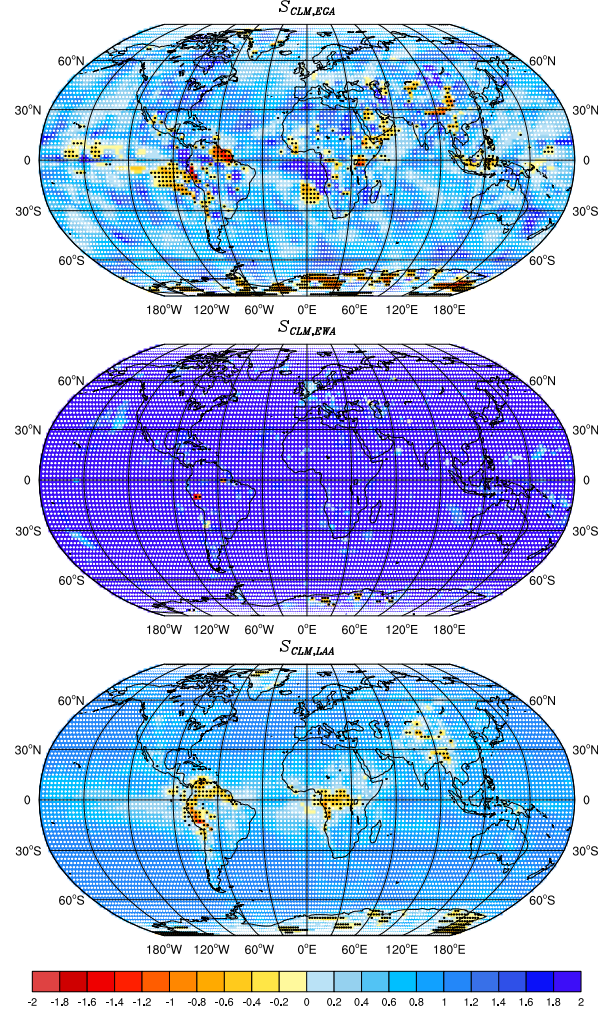


FIG. 11: Spatial distribution of the surface meridional wind  $STD$  skill score. Upper panel: EGA, middle panel: EWA, and lower panel: LAA. Positive values correspond to a smaller  $STD$  than the equally weighted ensemble and vice versa. White circles represent a statistically significant reduction of the  $STD$  and black circles represent a statistically significant increase of the  $STD$  relative to the  $STD$  of the equally weighted ensemble.

$STD$  (resulting in an  $S_{AVR,EWA}$  around 2 over the entire globe). The LAA also converges to the climatology, but due to the built-in switching probability, the weight assigned to the climatology is slightly smaller than 1, and accordingly, the  $S_{AVR,LAA}$  is slightly smaller than 2.



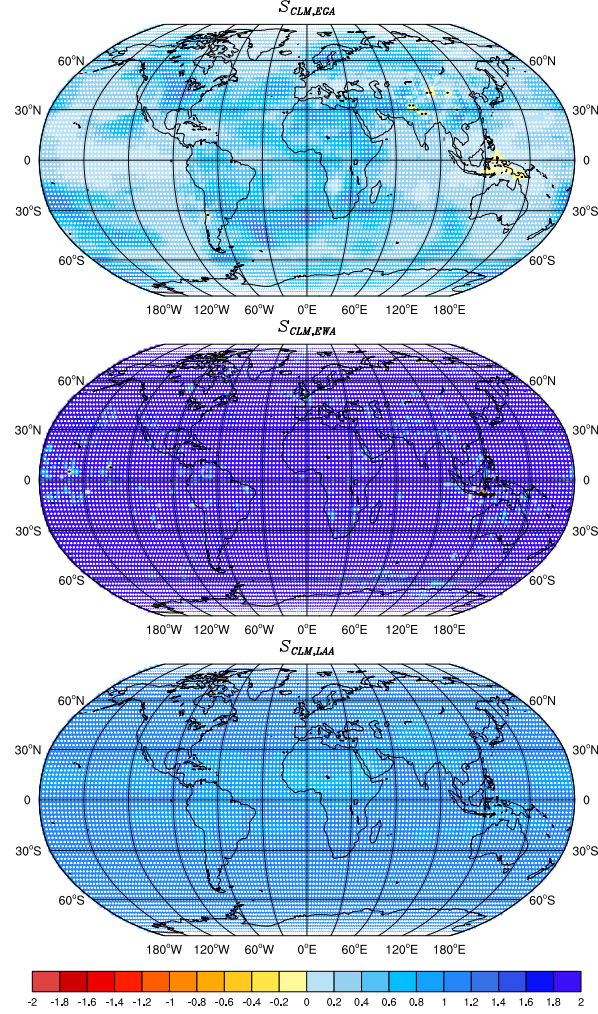


FIG. 12: Spatial distribution of the surface pressure  $STD$  skill score. Upper panel: EGA, middle panel: EWA, and lower panel: LAA. Positive values correspond to a smaller  $STD$  than the equally weighted ensemble and vice versa. White circles represent a statistically significant reduction of the  $STD$  and black circles represent a statistically significant increase of the  $STD$  relative to the  $STD$  of the equally weighted ensemble.

## VII. EGA WEIGHTS

Some of the results above regarding the skill of the *forecasters* were explained by the weights assigned to the climatology. Due to its superior performance, compared with each of the models in the ensemble, it is expected that the SLAs would assign it a high weight. However, assigning too high a weight to the climatology implies that the *forecaster* is not capable of capturing deviations from the climatology due to the physical processes captured in the models. Ideally, *forecasters*

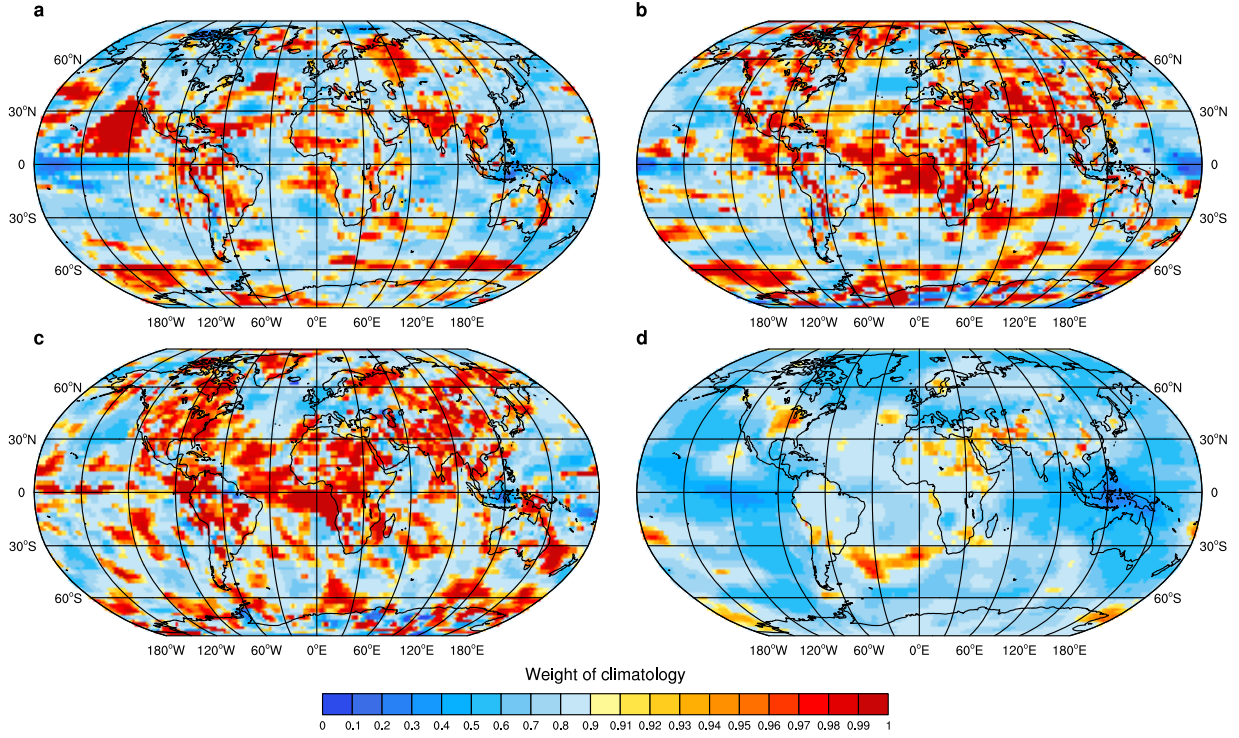


FIG. 13: Spatial distribution of the weight assigned to the climatology by the EGA *forecaster* for (a) surface temperature, (b) surface zonal wind, (c) surface meridional wind and (d) surface pressure.

should balance between the smaller  $RMSE$  of the climatology and the additional information available from the other models.

Figures 13, 14 and 15 show the spatial distribution of the weight assigned to the climatology, for each of the four climate variables, by the EGA, EWA and LAA, respectively. The weights in these figures correspond to the weights assigned at the end of the 20-year learning period (i.e., the weights used for the predictions). The colorbar was set to emphasize the differences. The EWA assigns the climatology weights close to 1 over the entire globe for the surface wind components and pressure. For the surface temperature, there are large regions in the tropics, close to the North Pole and along the coast of Antarctica where the weight of the climatology is not the only dominant *expert*. Similar patterns are observed for the LAA; however, the weight assigned to the climatology here is never 1 because this SLA is based on the fixed-share SLA that is designed to have a finite switching probability. Both the weights assigned by the EWA and those assigned by the LAA stem from the fact that these SLAs are designed to track the best *expert*, which in our ensemble turns out to be the climatology over most of the globe.

The EGA assigns a lower weight than the EWA and LAA to the climatology over most of the



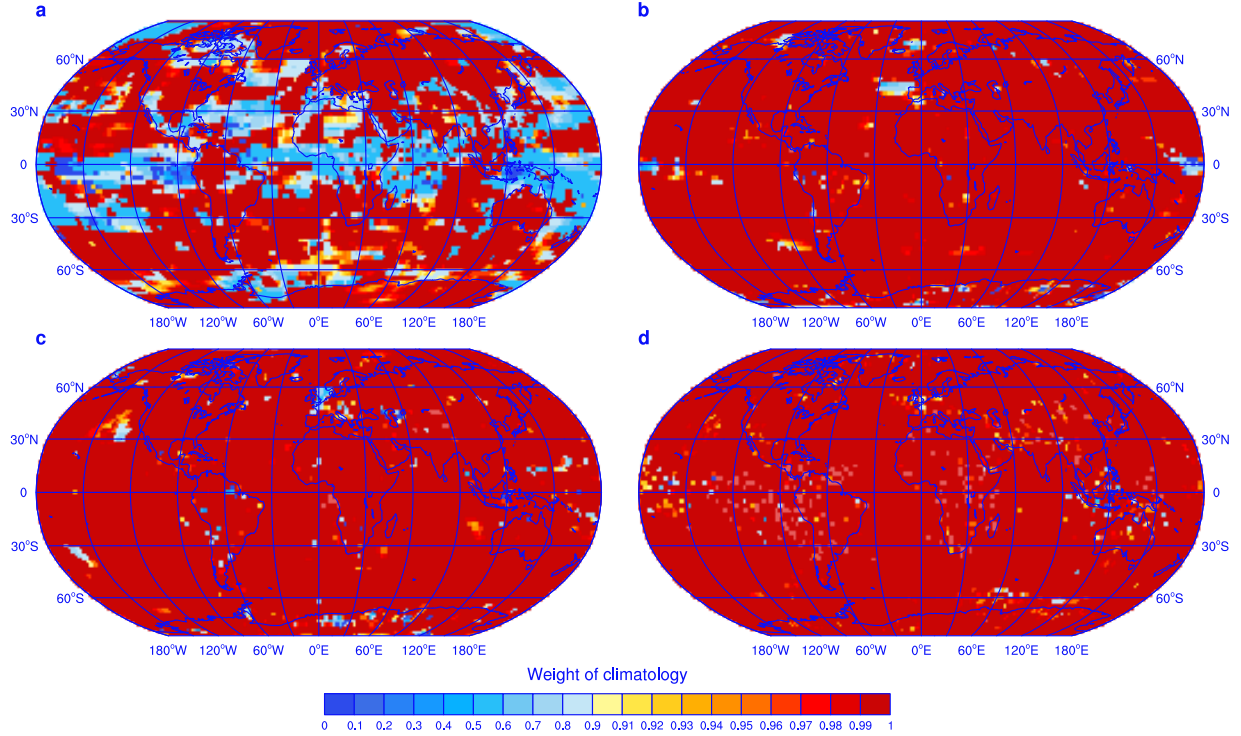


FIG. 14: Spatial distribution of the weight assigned to the climatology by the EWA *forecaster* for (a) surface temperature, (b) surface zonal wind, (c) surface meridional wind and (d) surface pressure.

globe for all the climate variables considered here. For the surface temperature, only in some regions (mostly in the eastern Pacific Ocean) are the predictions of the EGA dominated by the climatology. For the surface wind components, the regions dominated by the climatology are somewhat larger. The weight assigned to the climatology by the EGA for the surface pressure shows a much larger variability (note the nontrivial color map) than the weight assigned for the other variables. This variability also resulted in a somewhat poorer performance by the EGA in the predictions of this variable. This different performance for the surface pressure may be related to the lower quality of the data for this variable. Unlike the EWA and the LAA, the EGA is not designed to track the best *expert* but rather to track the measurements. Therefore, the lower weight assigned to the climatology suggests that useful information can be extracted from the models, and their ability to capture some of the processes affecting the climate dynamics in decadal time scales can be quantified by the weight assigned to them by the EGA.

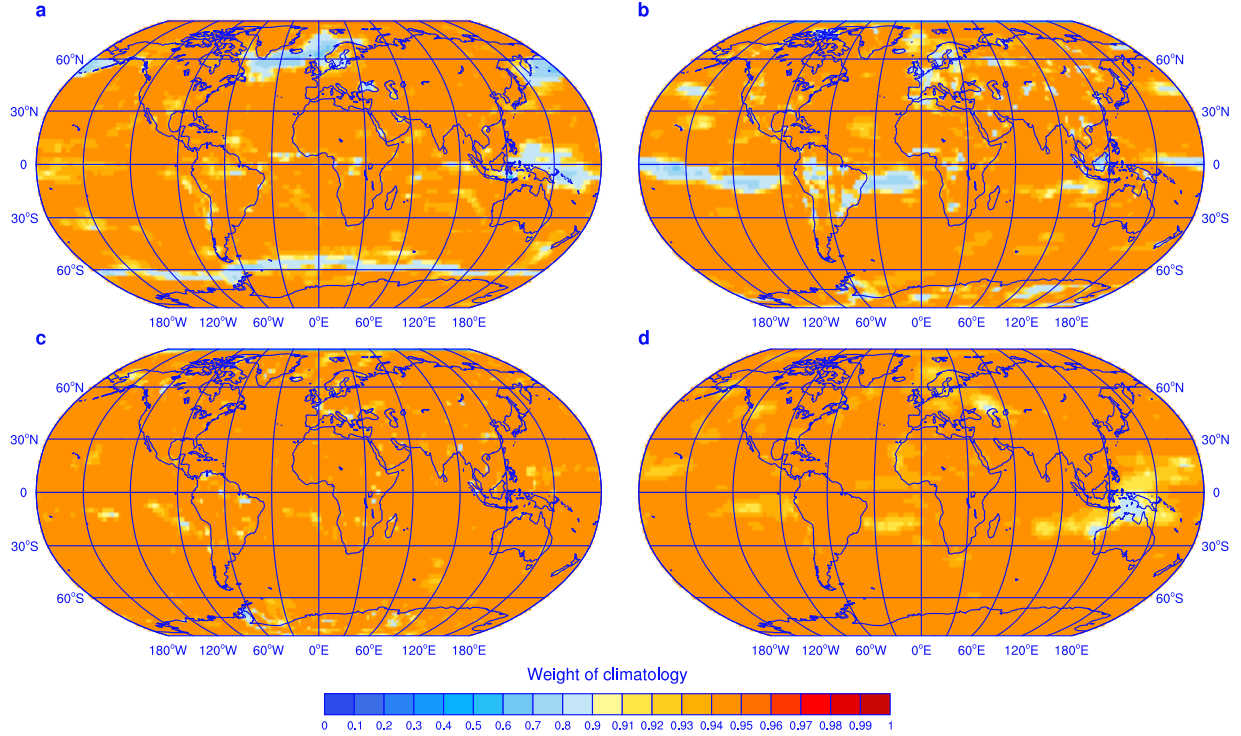


FIG. 15: Spatial distribution of the weight assigned to the climatology by the LAA *forecaster* for (a) surface temperature, (b) surface zonal wind, (c) surface meridional wind and (d) surface pressure.

## VIII. SUMMARY AND DISCUSSION

An ensemble of climate models is known to improve climate predictions and to help better assess the uncertainties associated with them. In this paper, we tested five different methods to combine the results of the decadal predictions of different models—EWA, EGA, LAA, REG and the equally weighted ensemble. The first three *forecasters* represent learning algorithms that weight the ensemble models according to their performances during a learning period. The REG attempts to find the linear combination of the model predictions that minimizes the sum of squared errors during the learning period, and the equally weighted ensemble represents no learning. We tried different learning periods and found the 20-year learning experiment to be the most promising. This learning period ensures that the learning exceeds well beyond the drift of the models. The *RMSE* and *STD* are smaller than those of shorter learning periods, and the results suggest that the lead time (the time from the initialization of the models) has a small effect. The predictions of the surface temperature, wind and pressure were studied, and their qualities were assessed.

The simple average was shown to have larger errors and larger uncertainties than the *forecasters*

that used a learning period to weight/combine the model predictions. The linear regression showed smaller errors than the equally weighted average. When no bias correction was applied to the data and the ensemble did not include the climatology, the errors of the regression were even smaller than those of the learning algorithms. However, in the more relevant ensemble that includes the climatology, the errors of the linear regression were higher than those of the learning algorithms. This poorer performance is associated with the basic assumptions of the linear regression and its oversimplified method to linearly combine the model predictions. The SLAs do not rely on these assumptions and use more advanced methods to weight the models, resulting in smaller errors. The REG method does not weight the models but rather finds an optimal linear combination of them; therefore, there is no straightforward method to estimate the uncertainties associated with the linear regression predictions. The EWA and the LAA were found to be more appropriate in cases in which tracking of the best model is of interest. The climatology outperformed all the other models; therefore, the EWA and the LAA converged to it over most of the globe and for all the four climate variables. Tracking the best model (by the EWA and LAA) was shown to result in too small uncertainties and thus in overconfident predictions. For the purpose of improving decadal climate predictions, we found the EGA to be more appropriate because it showed both the ability to reduce the errors and to provide more meaningful estimates of the uncertainties.

Although the globally averaged *RMSE* of the EGA is only a few percentage points smaller than that of the climatology, it was shown to be statistically significant. In addition, we found that in many regions, the improvement is larger. The spatial distribution of the EGA performance showed that it is skillful over large continuous regions. This finding suggests that the models were able to capture some physical processes that resulted in deviations from the climatology and that the EGA enabled the extraction of this additional information. Similarly, the large regions over which the climatology outperforms the *forecasters* may suggest that physical processes, associated with the climate dynamics affecting these regions, are not well captured by the models. The EGA performance was much poorer for the surface pressure than for the other variables. This poorer performance might be related to the quality of the models output or to the large fluctuations of this variable. The better predictions of the EWA and LAA for the surface pressure cannot be considered significant because their performance is similar to that of the climatology. The reduction of the uncertainties is much more substantial than the reduction of the errors and can reach to about 60 – 70%, globally. The uncertainties considered here are only those associated with the model variability within the ensemble. The internal uncertainties, scenario uncertainties

and other sources of uncertainty were not studied here.

The results presented here are in agreement with previous results (see Ref. [4] and references therein). However, in this work, monthly means were considered, whereas in previous works, the averages of longer periods, which have smaller fluctuations, were considered. A predictive skill of the EGA can be observed in the North Atlantic, in the North Indian Ocean and in some regions in the Pacific Ocean. In addition, the EGA showed predictive skill over many land areas, such as North Euro-Asia, Greenland, and, to some extent, also the Americas. The results suggest that learning algorithms can be used to improve climate predictions and to reduce the uncertainties associated with them.

### Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant number [293825]. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for the CMIP, and we thank the climate modeling groups (listed in Table I of this paper) for producing and making available their model output. For the CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. E.S. wishes to acknowledge a fellowship from the Israel Water Authority.

- 
- [1] Richard Moss, Mustafa Babiker, Sander Brinkman, Eduardo Calvo, Tim Carter, Jae Edmonds, Ismail Elgizouli, Seita Emori, Lin Erda, Kathy Hibbard, Roger Jones, Mikiko Kainuma, Jessica Kelleher, Jean Francois Lamarque, Martin Manning, Ben Matthews, Jerry Meehl, Leo Meyer, John Mitchell, Nebojsa Nakicenovic, Brian O'Neill, Ramon Pichs, Keywan Riahi, Steven Rose, Jean Pascal van Ypersel, Monika Zurek, Paul Runci, Ron Stouffer, Detlef van Vuuren, John Weyant, and Tom Wilbanks. Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Technical report, Geneva: Intergovernmental Panel on Climate Change, 2008.
- [2] M. Collins, R. Knutti, J. Arblaster, J. Dufresne, T. Fichet, P. Friedlingstein, X. Gao, W. Gutowski, T. Johns, G. Krinner, et al. Long-term climate change: Projections, commitments and irreversibility.

- In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1029–1136. Cambridge Univ. Press, Cambridge, UK, and New York, 2013.
- [3] M. Collins. Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):1957–1970, 2007.
  - [4] G. A. Meehl, L. Goddard, J. Murphy, R. J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. A. Giorgetta, A. M. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, and T. Stockdale. Decadal prediction. *Bulletin of the American Meteorological Society*, 90(10):1467–1485, 2009.
  - [5] Gerald A. Meehl, Lisa Goddard, George Boer, Robert Burgman, Grant Branstator, Christophe Cassou, Susanna Corti, Gokhan Danabasoglu, Francisco Doblas-Reyes, Ed Hawkins, Alicia Karspeck, Masahide Kimoto, Arun Kumar, Daniela Matei, Juliette Mignot, Rym Msadek, Antonio Navarra, Holger Pohlmann, Michele Rienecker, Tony Rosati, Edwin Schneider, Doug Smith, Rowan Sutton, Haiyan Teng, Geert Jan van Oldenborgh, Gabriel Vecchi, and Stephen Yeager. Decadal climate prediction: An update from the trenches. *Bulletin of the American Meteorological Society*, 95(2):243–267, 2013.
  - [6] T. T. Warner. *Numerical Weather and Climate Prediction*. Cambridge University Press, Cambridge, UK, 2011.
  - [7] M. A. Cane. Climate science: Decadal predictions in demand. *Nature Geoscience*, 3(4):231–232, 2010.
  - [8] D. M. Smith, S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317(5839):796–799, 2007.
  - [9] N. S. Keenlyside, M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner. Advancing decadal-scale climate prediction in the north atlantic sector. *Nature*, 453(7191):84–88, 2008.
  - [10] H. Pohlmann, J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke. Initializing decadal climate predictions with the gecco oceanic synthesis: Effects on the north atlantic. *Journal of Climate*, 22(14):3926–3938, 2009.
  - [11] F. J. Doblas-Reyes, M. Déqué, and J.-P. Piedelievre. Multi-model spread and probabilistic seasonal

- forecasts in provost. *Quarterly Journal of the Royal Meteorological Society*, 126(567):2069–2087, 2000.
- [12] F. J. Doblas-Reyes, V. Pavan, and D. B. Stephenson. The skill of multi-model seasonal forecasts of the wintertime north atlantic oscillation. *Climate Dynamics*, 21(5-6):501–514, 2003.
  - [13] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting—I. basic concept. *Tellus A*, 57(3):219–233, 2005.
  - [14] T. N. Palmer, F. J. Doblas-Reyes, R. Hagedorn, A. Alessandri, S. Gualdi, U. Andersen, H. Feddersen, P. Cantelaube, J-M. Terres, M. Davey, R. Graham, P. Délecluse, A. Lazar, M. Déqué, J-F. Guérémy, E. Díez, B. Orfila, M. Hoshen, A. P. Morse, N. Keenlyside, M. Latif, E. Maisonnavé, P. Rogel, V. Marletto, and M. C. Thomson. Development of a european multimodel ensemble system for seasonal-to-interannual prediction (demeter). *Bulletin of the American Meteorological Society*, 85(6):853–872, 2004.
  - [15] T. N. Palmer, Č Branković, and D. S. Richardson. A probability and decision-model analysis of provost seasonal multi-model ensemble integrations. *Quarterly Journal of the Royal Meteorological Society*, 126(567):2013–2033, 2000.
  - [16] H.-M. Kim, P. J. Webster, and J. A. Curry. Evaluation of short-term climate change prediction in multi-model cmip5 decadal hindcasts. *Geophysical Research Letters*, 39(10):L10701, 2012.
  - [17] E. Hawkins and R. Sutton. The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8):1095–1107, 2009.
  - [18] E. Strobach and G. Bel. The contribution of internal and model variabilities to the uncertainty in cmip5 decadal climate predictions, 2015.
  - [19] P. Cox and D. Stephenson. A changing climate for prediction. *Science*, 317(5835):207–208, 2007.
  - [20] Christoph M. Buser, H. R. Künsch, D. Lüthi, M. Wild, and C. Schär. Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dynamics*, 33(6):849–868, 2009.
  - [21] C. M. Buser, H. R. Künsch, and C. Schär. Bayesian multi-model projections of climate: generalization and application to ENSEMBLES results. *Climate Research*, 44(2-3):227–241, 2010.
  - [22] R. L. Smith, C. Tebaldi, D. Nychka, and L. O. Mearns. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104(485):97–116, 2009.
  - [23] C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns. Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540, 2005.

- [24] C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, 2007.
- [25] R. Furrer, S. R. Sain, D. Nychka, and G. A. Meehl. Multivariate bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics*, 14(3):249–266, 2007.
- [26] A. M. Greene, L. Goddard, and U. Lall. Probabilistic multimodel regional temperature change projections. *Journal of Climate*, 19(17):4326–4343, 2006.
- [27] J. M. Murphy, D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001):768–772, 2004.
- [28] J. Räisänen, L. Ruokolainen, and J. Ylhäisi. Weighting of model results for improving best estimates of climate change. *Climate Dynamics*, 35(2-3):407–422, 2010.
- [29] B. Rajagopalan, U. Lall, and S. E. Zebiak. Categorical climate forecasts through regularization and optimal combination of multiple gcm ensembles. *Monthly Weather Review*, 130(7):1792, 2002.
- [30] A. W. Robertson, U. Lall, S. E. Zebiak, and L. Goddard. Improved combination of multiple atmospheric gcm ensembles for seasonal prediction. *Monthly Weather Review*, 132(12):2732–2744, 2004.
- [31] J. Feng, D.-K. Lee, C. Fu, J. Tang, Y. Sato, H. Kato, J. L. McGregor, and K. Mabuchi. Comparison of four ensemble methods combining regional climate simulations over asia. *Meteorology and Atmospheric Physics*, 111(1-2):41–53, 2011.
- [32] A. Chakraborty and T. N. Krishnamurti. Improving global model precipitation forecasts over india using downscaling and the fsu superensemble. part ii: Seasonal climate. *Monthly Weather Review*, 137(9):2736–2757, 2009.
- [33] F. J. Doblas-Reyes, R. Hagedorn, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting–II. calibration and combination. *Tellus A*, 57(3):234–252, 2005.
- [34] K. Fraedrich and N. R. Smith. Combining predictive schemes in long-range forecasting. *Journal of Climate*, 2(3):291–294, 1989.
- [35] V. V. Kharin and F. W. Zwiers. Climate predictions with multimodel ensembles. *Journal of Climate*, 15(7):793–799, 2002.
- [36] T. N. Krishnamurti. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550, 1999.
- [37] T. N. Krishnamurti, C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and

- S. Surendran. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13(23):4196–4216, 2000.
- [38] V. Pavan and F. J. Doblas-Reyes. Multi-model seasonal hindcasts over the euro-atlantic: skill scores and dynamic features. *Climate dynamics*, 16(8):611–625, 2000.
- [39] M. Peña and H. van den Dool. Consolidation of multimodel forecasts by ridge regression: Application to pacific sea surface temperature. *Journal of Climate*, 21(24):6521–6538, 2008.
- [40] P. Peng, A. Kumar, H. van den Dool, and A. G. Barnston. An analysis of multimodel ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research: Atmospheres*, 107(D23):ACL 18–1–ACL 18–12, 2002.
- [41] W. T. Yun, L. Stefanova, A. K. Mitra, T. S. V. Vijaya Kumar, W. Dewar, and T. N. Krishnamurti. A multi-model superensemble algorithm for seasonal climate prediction using demeter forecasts. *Tellus A*, 57(3):280–289, 2005.
- [42] W. T. Yun, L. Stefanova, and T. N. Krishnamurti. Improvement of the multimodel superensemble technique for seasonal forecasts. *Journal of climate*, 16(22):3834–3840, 2003.
- [43] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, UK, 2006.
- [44] C. Monteleoni, G. Schmidt, and S. Saroha. Tracking climate model. In *NASA Conference on Intelligent Data Understanding (CIDU)*, pages 1–15, 2010.
- [45] C. Monteleoni, G. A. Schmidt, S. Saroha, and E. Asplund. Tracking climate models. *Statistical Analysis and Data Mining*, 4(4):372–392, 2011.
- [46] V. Mallet, G. Stoltz, and B. Mauricette. Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research: Atmospheres*, 114(D5):D050307, 2009.
- [47] Vivien Mallet. Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research: Atmospheres*, 115(D24):D24303, 2010.
- [48] E. Strobach and G. Bel. Improvement of climate predictions and reduction of their uncertainties using learning algorithms. *Atmospheric Chemistry and Physics*, 15:8631–8641, 2015.
- [49] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. A Summary of the CMIP5 Experiment Design. [http://cmip-pcmdi.llnl.gov/cmip5/experiment\\_design.html](http://cmip-pcmdi.llnl.gov/cmip5/experiment_design.html), 2009. Accessed: 2015-07-11.
- [50] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-year reanal-



- ysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.
- [51] B. Kirtman, S. B. Power, J. A. Adedoyin, G. J. Boer, R. Bojariu, I. Camilloni, F. J. Doblas-Reyes, A. M. Fiore, M. Kimoto, G. A. Meehl, M. Prather, A. Sarr, C. Schär, R. Sutton, G. J. van Oldenborgh, G. Vecchi, and H. J. Wang. Near-term climate change: Projections and predictability. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 953–1028. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.
  - [52] F. J. Doblas-Reyes, I. Andreu-Burillo, Y. Chikamoto, J. Garcia-Serrano, V. Guemas, M. Kimoto, T. Mochizuki, L. R. L. Rodrigues, and G. J. van Oldenborgh. Initialized near-term regional climate change prediction. *Nature Communications*, 4:1715, 2013.
  - [53] W. A. Müller, J. Baehr, H. Haak, J. H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J. S. von Storch, and J. Marotzke. Forecast skill of multi-year seasonal means in the decadal prediction system of the max planck institute for meteorology. *Geophysical Research Letters*, 39(22):L22707, 2012.
  - [54] W. A. Müller, H. Pohlmann, F. Sienz, and D. Smith. Decadal climate predictions for the period 1901–2010 with a coupled climate model. *Geophysical Research Letters*, 41(6):2100–2107, 2014.
  - [55] T. Kruschke, H. Rust, C. Kadow, G. Leckebusch, and U. Ulbrich. Evaluating decadal predictions of northern hemispheric cyclone frequencies. *Tellus A*, 66(0), 2014.
  - [56] S. M. Uppala, P. W. Kållberg, A. J. Simmons, U. Andrae, V. Da Costa Bechtold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. Van De Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, R. Jenne, A. P. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo, and J. Woollen. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012, 2005.
  - [57] Kazutoshi Onogi, Junichi TsuTsui, Hiroshi Koide, Masami Sakamoto, Shinya Kobayashi, Hiroaki Hatsushika, Takanori Matsumoto, Nobuo Yamazaki, Hirotaka Kamahori, Kiyotoshi Takahashi, Shinji Kadokura, Koji Wada, Koji Kato, Ryo Oyama, Tomoaki Ose, Nobutaka Mannoji, and Ryusuke Taira. The jra-25 reanalysis. *Journal of the Meteorological Society of Japan. Ser. II*, 85(3):369–432, 2007.
  - [58] L. Goddard, A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield,

- C. Deser, S.J. Mason, B.P. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C.A.T. Ferro, D.B. Stephenson, G.A. Meehl, T. Stockdale, R. Burgman, A.M. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, and T. Delworth. A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, 40(1-2):245–272, 2013.
- [59] C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *Advances in Neural Information Processing Systems*, volume 16, pages 1093–1100. 2003.
- [60] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

## **Appendix A: Supplementary Information**

This Supplementary Information provides the globally averaged root mean square errors for the different *forecasters* and different learning periods. The results provided here were used to select the optimal bias correction method for each *forecaster* and each climate variable. In addition, the globally averaged standard deviations of the ensemble weighted by the different *forecasters* are provided.

TABLE III: The surface temperature  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.4432	1.4267	1.4177	1.3716	1.3876
	Avg. correction	1.3624	1.3394	1.346	1.2972	1.3143
	Clm. correction	1.3257	1.2584	1.2614	1.2123	1.2312
EWA	No correction	1.4812	1.4687	1.4755	1.4287	1.4552
	Avg. correction	1.3811	1.356	1.3573	1.3158	1.3349
	Clm. correction	1.3397	1.2632	1.2556	1.2086	1.2317
LAA	No correction	1.4764	1.4898	1.5233	1.5062	1.5472
	Avg. correction	1.3595	1.368	1.3948	1.3717	1.4013
	Clm. correction	1.3193	1.2825	1.2906	1.2593	1.288
REG	No correction	1.4895	1.3632	1.3376	1.2922	1.301
	Avg. correction	1.4789	1.3607	1.3381	1.2902	1.2989
	Clm. correction	1.5446	1.3241	1.2868	1.2304	1.2415
AVG	No correction	1.7776	1.7918	1.8111	1.8046	1.8422
	Avg. correction	1.4152	1.4118	1.4184	1.3904	1.4194
	Clm. correction	1.2898	1.2475	1.2472	1.2038	1.2338
CLM		1.2393	1.1994	1.2213	1.1881	1.1956

TABLE IV: The zonal surface wind  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.7814	1.7694	1.7765	1.7618	1.8004
	Avg. correction	1.6905	1.6806	1.6829	1.6698	1.7076
	Clm. correction	1.7468	1.672	1.6589	1.628	1.667
EWA	No correction	1.8047	1.7979	1.8027	1.7946	1.8312
	Avg. correction	1.7021	1.6934	1.695	1.6806	1.7221
	Clm. correction	1.7589	1.6802	1.6634	1.6297	1.6738
LAA	No correction	1.8114	1.8259	1.8557	1.8678	1.9199
	Avg. correction	1.7105	1.7234	1.7499	1.75	1.8074
	Clm. correction	1.758	1.7014	1.7077	1.6873	1.7431
REG	No correction	1.7633	1.7197	1.7134	1.699	1.7366
	Avg. correction	1.7296	1.6865	1.6801	1.6637	1.7034
	Clm. correction	1.8124	1.6892	1.6645	1.6277	1.6662
AVG	No correction	1.8947	1.8982	1.9094	1.9126	1.9665
	Avg. correction	1.7182	1.7112	1.7188	1.7072	1.7519
	Clm. correction	1.7429	1.6697	1.6629	1.6312	1.677
CLM		1.6285	1.5719	1.569	1.5323	1.5692

TABLE V: The meridional surface wind  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.4532	1.4502	1.4575	1.4488	1.4701
	Avg. correction	1.3923	1.3839	1.3897	1.3808	1.3987
	Clm. correction	1.4251	1.3607	1.3553	1.3336	1.3474
EWA	No correction	1.4722	1.4703	1.4765	1.4698	1.4937
	Avg. correction	1.4033	1.3941	1.398	1.3892	1.4102
	Clm. correction	1.4355	1.3647	1.3577	1.3355	1.3528
LAA	No correction	1.4756	1.493	1.5169	1.5275	1.5599
	Avg. correction	1.4065	1.4162	1.4371	1.4412	1.47
	Clm. correction	1.4336	1.3829	1.3896	1.3792	1.4019
REG	No correction	1.4412	1.4078	1.4081	1.3956	1.4149
	Avg. correction	1.4171	1.3839	1.3831	1.3692	1.3867
	Clm. correction	1.4673	1.3651	1.3526	1.3279	1.3401
AVG	No correction	1.5319	1.5349	1.5424	1.544	1.5708
	Avg. correction	1.4145	1.4109	1.4174	1.4114	1.4323
	Clm. correction	1.4185	1.36	1.3563	1.3355	1.352
CLM		1.3352	1.2853	1.2876	1.2617	1.2731

TABLE VI: The surface pressure  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	3.7837	3.735	3.7369	3.7327	3.8133
	Avg. correction	2.6627	2.6388	2.6592	2.6314	2.7475
	Clm. correction	2.7774	2.628	2.6278	2.5765	2.682
EWA	No correction	3.9589	3.9483	3.9562	3.9528	4.043
	Avg. correction	2.689	2.6561	2.6687	2.6458	2.7556
	Clm. correction	2.8111	2.6444	2.6335	2.5824	2.6838
LAA	No correction	4.1878	4.1653	4.2056	4.2313	4.3497
	Avg. correction	2.7081	2.7503	2.8186	2.8327	2.9609
	Clm. correction	2.8254	2.7417	2.7802	2.7597	2.8885
REG	No correction	2.8237	2.6976	2.6875	2.6583	2.7572
	Avg. correction	2.824	2.6978	2.6877	2.6584	2.7574
	Clm. correction	3.035	2.7184	2.6809	2.6158	2.7049
AVG	No correction	6.4757	6.4749	6.4933	6.4998	6.5855
	Avg. correction	2.6786	2.659	2.6839	2.6639	2.7813
	Clm. correction	2.7506	2.6155	2.6208	2.5701	2.6785
CLM		2.5746	2.4531	2.4606	2.3935	2.4933

TABLE VII: The surface temperature  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.229	1.1919	1.2146	1.1562	1.1516
	Avg. correction	1.2204	1.1909	1.2088	1.1681	1.17
	Clm. correction	1.2616	1.2095	1.2202	1.177	1.183
EWA	No correction	1.2382	1.1992	1.2207	1.1869	1.1937
	Avg. correction	1.2389	1.1993	1.2205	1.1873	1.1926
	Clm. correction	1.2468	1.1993	1.2157	1.1797	1.1825
LAA	No correction	1.2179	1.1907	1.2141	1.1825	1.1931
	Avg. correction	1.2125	1.1865	1.21	1.1787	1.1888
	Clm. correction	1.2354	1.1945	1.2146	1.1814	1.1904
REG	No correction	1.3822	1.2546	1.2539	1.2035	1.1928
	Avg. correction	1.3784	1.2526	1.2537	1.2046	1.1918
	Clm. correction	1.4939	1.291	1.2645	1.2073	1.1915
AVG	No correction	1.6645	1.6783	1.6983	1.6896	1.726
	Avg. correction	1.3597	1.3562	1.3644	1.3366	1.3649
	Clm. correction	1.2748	1.2328	1.2342	1.1913	1.219
CLM		1.2393	1.1994	1.2213	1.1881	1.1956

TABLE VIII: The zonal surface wind  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.6029	1.5622	1.565	1.5288	1.5642
	Avg. correction	1.6004	1.5643	1.5636	1.5365	1.5722
	Clm. correction	1.6666	1.5962	1.585	1.5525	1.5881
EWA	No correction	1.6264	1.5708	1.5686	1.531	1.5684
	Avg. correction	1.6251	1.5712	1.5685	1.5324	1.57
	Clm. correction	1.6382	1.5741	1.57	1.5347	1.5718
LAA	No correction	1.6016	1.5622	1.5634	1.5305	1.569
	Avg. correction	1.6026	1.5629	1.5636	1.53	1.5679
	Clm. correction	1.6356	1.5732	1.5693	1.5332	1.5708
REG	No correction	1.682	1.6037	1.5887	1.5486	1.5819
	Avg. correction	1.6844	1.605	1.5892	1.5493	1.5832
	Clm. correction	1.7564	1.6348	1.6009	1.5623	1.5911
AVG	No correction	1.8141	1.8157	1.8275	1.8284	1.8813
	Avg. correction	1.6765	1.6683	1.6758	1.6631	1.7073
	Clm. correction	1.7176	1.6471	1.641	1.6092	1.6541
CLM		1.6285	1.5719	1.569	1.5323	1.5692



TABLE IX: The meridional surface wind  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.3119	1.2777	1.282	1.2587	1.2707
	Avg. correction	1.3108	1.2785	1.2835	1.263	1.2748
	Clm. correction	1.364	1.3025	1.3002	1.2769	1.287
EWA	No correction	1.3339	1.2848	1.2874	1.2607	1.2737
	Avg. correction	1.3329	1.2846	1.2872	1.2616	1.2731
	Clm. correction	1.3411	1.2874	1.2885	1.2627	1.2742
LAA	No correction	1.3102	1.2762	1.2816	1.2588	1.2716
	Avg. correction	1.3124	1.2779	1.2829	1.2594	1.2717
	Clm. correction	1.3395	1.286	1.2875	1.2621	1.2738
REG	No correction	1.3744	1.3068	1.3023	1.274	1.2851
	Avg. correction	1.3753	1.3075	1.3028	1.2754	1.2861
	Clm. correction	1.4337	1.3232	1.3111	1.2837	1.2929
AVG	No correction	1.4671	1.4692	1.4776	1.4779	1.5038
	Avg. correction	1.3774	1.3732	1.3804	1.3735	1.3935
	Clm. correction	1.3996	1.3427	1.34	1.3188	1.3347
CLM		1.3352	1.2853	1.2876	1.2617	1.2731

TABLE X: The surface pressure  $RMSE_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	2.5682	2.465	2.4716	2.4189	2.4999
	Avg. correction	2.5228	2.4518	2.4667	2.412	2.5131
	Clm. correction	2.6355	2.4967	2.4986	2.4404	2.5374
EWA	No correction	2.5735	2.453	2.4607	2.3942	2.4938
	Avg. correction	2.5712	2.4528	2.4603	2.3942	2.4942
	Clm. correction	2.5833	2.4569	2.4631	2.3978	2.4976
LAA	No correction	2.6987	2.5417	2.5194	2.4457	2.536
	Avg. correction	2.5318	2.4412	2.454	2.3912	2.4932
	Clm. correction	2.5821	2.4552	2.4614	2.3951	2.4957
REG	No correction	2.7339	2.5263	2.5076	2.4396	2.5213
	Avg. correction	2.7338	2.5263	2.5076	2.4396	2.5213
	Clm. correction	2.8891	2.5684	2.5303	2.4565	2.5325
AVG	No correction	5.9201	5.917	5.9363	5.9402	6.0273
	Avg. correction	2.6251	2.6026	2.6277	2.6045	2.7222
	Clm. correction	2.7146	2.5816	2.5876	2.5354	2.6441
CLM		2.5746	2.4531	2.4606	2.3935	2.4933

TABLE XI: The surface temperature  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.5722	1.5516	1.5509	1.5003	1.4699
	Avg. correction	1.2071	1.1681	1.1419	1.1	1.0898
	Clm. correction	1.1107	1.047	1.0066	0.97441	0.96212
EWA	No correction	1.1803	1.1255	1.0967	1.0734	1.0492
	Avg. correction	1.0505	1.0143	0.99385	0.96995	0.96079
	Clm. correction	1.0001	0.9584	0.9334	0.9094	0.90286
LAA	No correction	1.3797	1.1943	1.0777	0.9866	0.91267
	Avg. correction	1.1392	1.0059	0.91539	0.83884	0.77893
	Clm. correction	1.0596	0.92525	0.83163	0.77086	0.71868
AVR	No correction	1.7773	1.7817	1.7849	1.7888	1.7888
	Year bias	1.2573	1.2335	1.2129	1.2	1.1874
	Month bias	1.1132	1.053	1.0151	0.99245	0.9729

TABLE XII: The zonal surface wind  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.3211	1.3083	1.2938	1.2958	1.301
	Avg. correction	1.218	1.1991	1.1895	1.1955	1.1906
	Clm. correction	1.1857	1.1267	1.1117	1.1121	1.1013
EWA	No correction	1.1648	1.1222	1.1094	1.1066	1.1125
	Avg. correction	1.1275	1.0924	1.0866	1.0938	1.0903
	Clm. correction	1.0902	1.034	1.0298	1.0247	1.0193
LAA	No correction	1.2084	1.0821	0.98004	0.91758	0.86497
	Avg. correction	1.1193	1.0126	0.92726	0.88286	0.83293
	Clm. correction	1.0989	0.96991	0.88354	0.83389	0.7855
AVR	No correction	1.3665	1.3673	1.3694	1.3751	1.377
	Avg. correction	1.1962	1.189	1.1858	1.1892	1.187
	Clm. correction	1.1566	1.1034	1.0871	1.085	1.0759

TABLE XIII: The meridional surface wind  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	1.0428	1.0292	1.0279	1.0221	1.0122
	Avg. correction	0.959	0.95016	0.94543	0.94398	0.93683
	Clm. correction	0.93743	0.9019	0.88682	0.88115	0.86831
EWA	No correction	0.91457	0.88655	0.87795	0.87873	0.86699
	Avg. correction	0.88297	0.87172	0.86361	0.86215	0.85405
	Clm. correction	0.85492	0.83007	0.81715	0.80731	0.79725
LAA	No correction	0.94936	0.85194	0.78035	0.7294	0.67994
	Avg. correction	0.88457	0.80834	0.74702	0.70583	0.66468
	Clm. correction	0.86813	0.76843	0.69957	0.65128	0.60696
AVR	No correction	1.0623	1.0622	1.0646	1.0653	1.0663
	Avg. correction	0.9454	0.93813	0.93673	0.93579	0.9352
	Clm. correction	0.92446	0.88248	0.86953	0.86301	0.85696

TABLE XIV: The surface pressure  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	4.4105	4.351	4.3096	4.2797	4.2831
	Avg. correction	2.4633	2.4556	2.451	2.453	2.4329
	Clm. correction	2.4548	2.3638	2.3503	2.343	2.314
EWA	No correction	2.5294	2.3962	2.3722	2.3379	2.327
	Avg. correction	2.3429	2.3491	2.3509	2.3472	2.3384
	Clm. correction	2.3384	2.3005	2.3073	2.2899	2.2663
LAA	No correction	3.7348	3.267	2.9498	2.7572	2.5945
	Avg. correction	2.2853	2.1309	2.0012	1.9073	1.8249
	Clm. correction	2.2975	2.0974	1.9856	1.8839	1.7998
AVR	No correction	5.8832	5.885	5.8841	5.8911	5.8787
	Avg. correction	2.4017	2.3929	2.3839	2.3834	2.3636
	Clm. correction	2.3771	2.2727	2.2459	2.2339	2.2015

TABLE XV: The surface temperature  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	0.85636	0.79338	0.74092	0.74888	0.77041
	Avg. correction	0.69405	0.65224	0.60268	0.60546	0.59947
	Clm. correction	0.82139	0.76932	0.72812	0.69949	0.68463
EWA	No correction	0.069617	0.04385	0.034291	0.034327	0.046811
	Avg. correction	0.093904	0.055604	0.042137	0.042974	0.0541
	Clm. correction	0.32004	0.22664	0.18245	0.18477	0.22204
LAA	No correction	0.86891	0.69606	0.60843	0.56257	0.52853
	Avg. correction	0.64118	0.49176	0.41823	0.37824	0.35085
	Clm. correction	0.61502	0.44528	0.35476	0.31043	0.28258
AVR	No correction	1.8474	1.8482	1.8513	1.8555	1.8556
	Avg. correction	1.301	1.2736	1.2536	1.2398	1.2267
	Clm. correction	1.1244	1.0643	1.0281	1.006	0.98709

TABLE XVI: The zonal surface wind  $STD_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	0.73276	0.65502	0.60472	0.6103	0.55895
	Avg. correction	0.73511	0.66343	0.60299	0.60858	0.55592
	Clm. correction	0.84851	0.76309	0.71594	0.69938	0.65361
EWA	No correction	0.084994	0.046667	0.035807	0.034519	0.030238
	Avg. correction	0.11876	0.066557	0.052488	0.046515	0.041827
	Clm. correction	0.34414	0.19312	0.13497	0.11729	0.10056
LAA	No correction	0.70173	0.54873	0.47246	0.43713	0.41112
	Avg. correction	0.60517	0.46204	0.39078	0.35748	0.33391
	Clm. correction	0.60982	0.43764	0.35328	0.3177	0.29133
AVR	No correction	1.439	1.4335	1.4333	1.4364	1.4383
	Avg. correction	1.2418	1.2278	1.2233	1.2237	1.2214
	Clm. correction	1.173	1.1189	1.1023	1.0993	1.0906



TABLE XVII: The meridional surface wind  $STD_{GAW}$  for the different *forecasters* and bias correction methods. The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	0.56695	0.50983	0.47015	0.47185	0.44052
	Avg. correction	0.56347	0.52115	0.47243	0.47293	0.43637
	Clm. correction	0.67315	0.61536	0.56948	0.55546	0.52289
EWA	No correction	0.056521	0.031311	0.025234	0.025545	0.022147
	Avg. correction	0.078621	0.046682	0.036763	0.030673	0.026266
	Clm. correction	0.26654	0.15412	0.11447	0.094416	0.080736
LAA	No correction	0.54687	0.42964	0.37097	0.3416	0.32204
	Avg. correction	0.48228	0.36769	0.31136	0.28407	0.26622
	Clm. correction	0.47982	0.34436	0.27938	0.25008	0.22985
AVR	No correction	1.1258	1.119	1.1194	1.1178	1.1187
	Avg. correction	0.98815	0.9738	0.97075	0.96766	0.96716
	Clm. correction	0.93731	0.89451	0.88127	0.87434	0.86855

TABLE XVIII: The surface pressure  $STD_{GAW}$  for the different *forecasters* and bias correction methods.

The climatology is included in the ensemble.

Forecaster	Bias correction	Learning period				
		5	10	15	20	25
EGA	No correction	2.9392	2.717	2.5768	2.5104	2.3833
	Avg. correction	1.5787	1.5048	1.4308	1.3989	1.2599
	Clm. correction	1.7647	1.631	1.5915	1.5789	1.4751
EWA	No correction	0.18324	0.099206	0.073743	0.064644	0.055238
	Avg. correction	0.19148	0.13322	0.10974	0.088419	0.07355
	Clm. correction	0.60822	0.39215	0.31246	0.27342	0.22431
LAA	No correction	2.9241	2.4551	2.1702	2.0292	1.9133
	Avg. correction	1.1266	0.85932	0.71971	0.65848	0.60953
	Clm. correction	1.1816	0.83857	0.67345	0.60363	0.54823
AVR	No correction	6.2919	6.2863	6.283	6.2882	6.2762
	Avg. correction	2.4483	2.4306	2.42	2.4176	2.3972
	Clm. correction	2.3963	2.2911	2.2639	2.2515	2.2187